



Universidade de Brasília

Instituto de Ciências Exatas  
Departamento de Ciência da Computação

# GraphRAG para Prontuários Eletrônicos: Uma Abordagem Aumentada por Grafos para Acesso à Informação em Saúde

João Vitor R. Baptista

Dissertação apresentada como requisito parcial para conclusão do  
Mestrado Profissional em Computação Aplicada

Orientador

Prof. Dr. Luís Paulo Faina Garcia

Brasília  
2026

Ficha catalográfica elaborada automaticamente,  
com os dados fornecidos pelo(a) autor(a)

B222g      Baptista, João Vitor Rodrigues  
              GraphRAG para Prontuários Eletrônicos: Uma Abordagem  
              Aumentada por Grafos para Acesso à Informação em Saúde /  
              João Vitor Rodrigues Baptista; orientador Luís Paulo Faina  
              Gargia. Brasília, 2026.  
              92 p.

              Dissertação(Mestrado Profissional em Computação Aplicada)  
              Universidade de Brasília, 2026.

              1. GenIA. 2. GraphRAG. 3. RAG. 4. LLM. 5. Grafos. I.  
              Gargia, Luís Paulo Faina, orient. II. Título.



Universidade de Brasília

Instituto de Ciências Exatas  
Departamento de Ciência da Computação

# **GraphRAG para Prontuários Eletrônicos: Uma Abordagem Aumentada por Grafos para Acesso à Informação em Saúde**

João Vitor R. Baptista

Dissertação apresentada como requisito parcial para conclusão do  
Mestrado Profissional em Computação Aplicada

Prof. Dr. Luís Paulo Faina Garcia (Orientador)  
CIC/UnB

Prof. Dr. Murillo Guimarães Carneiro    Prof. Dr. Thiago de Paulo Faleiros  
(Membro externo - FACOM/UFU)    (Membro Interno - CIC/UnB)

Prof. Dr. Glauco Vitor Pedrosa  
(Suplente - FCTE/UnB)

Prof.a Dr.a Edna Dias Canedo  
Coordenadora do Programa de Pós-graduação em Computação Aplicada

Brasília, 13 de Março de 2026

# Dedicatória

*Dedico esta, assim como todas as demais conquistas, a minha noiva Stefeny Oliveira de Medeiros, para a flor da minha vida Rosângela Rodrigues da Silva, ao meu amado pai, João dos Santos Baptista e a irmã que tanto amo Maria Victória Rodrigues Baptista*

# Agradecimentos

*A minha incrível família por todo suporte e carinho que possuem por mim. Principalmente, e minha noiva e futura esposa Stefeny Oliveira de Medeiros, a mulher que é luz da minha vida, minha querida mãe, Rosângela Rodrigues da Silva e meu amado pai, João dos Santos Baptista e a minha super irmã Maria Victória Rodrigues Baptista que me aturam todos os dias.*

# Resumo

A crescente importância da saúde digital e a necessidade de uma maior literacia em saúde exigem métodos eficazes para acessar e compreender os Registos Eletrônicos de Saúde (EHRs). Embora os Modelos de Linguagem de Grande Escala (LLMs) se mostrem promissores neste domínio, a Geração Aumentada por Recuperação (RAG) tradicional enfrenta dificuldades em lidar com a natureza complexa e interconectada dos dados clínicos. A Geração Aumentada por Recuperação baseada em Grafos (*GraphRAG*) emerge como uma alternativa poderosa, aproveitando grafos de conhecimento (KGs) para capturar as relações semânticas nos EHRs. Este trabalho avalia a eficácia da expansão de grafos numa arquitetura *GraphRAG* para melhorar a recuperação de informação a partir de dados médicos formatados em FHIR. Propõem-se uma abordagem de expansão de 1-hop construída sobre uma base de pesquisa por similaridade e lexical que, embora herde algumas limitações da recuperação tradicional baseada em palavras-chave e similaridade, melhora significativamente o acesso dos LLMs a informações contextuais abrangentes e diversificadas. A nossa avaliação, utilizando dados sintéticos de pacientes e um conjunto direcionado de questões em cinco LLMs distintos, revela que a estratégia de expansão de 1-hop supera consistentemente a abordagem base em métricas subjetivas, como abrangência e diversidade, e frequentemente em métricas qualitativas, como a relevância da resposta e do contexto.

**Palavras-chave:** *GraphRAG*, *RAG*, *LLM*, *FHIR*, Grafos, Saúde

# Abstract

The growing importance of digital health and the need for greater health literacy demand effective methods for accessing and understanding Electronic Health Records (EHRs). Although Large Language Models (LLMs) show promise in this domain, traditional Retrieval-Augmented Generation (RAG) struggles to handle the complex and interconnected nature of clinical data. GraphRAG emerges as a powerful alternative, leveraging knowledge graphs (KGs) to capture the semantic relationships within EHRs. This research evaluates the effectiveness of graph expansion in a GraphRAG architecture to improve information retrieval from FHIR-formatted medical data. We propose a 1-hop expansion approach built upon a similarity and lexical search baseline which, while inheriting some limitations of traditional keyword and similarity-based retrieval, significantly improves LLMs' access to comprehensive and diverse contextual information. Our evaluation, using synthetic patient data and a targeted set of questions across five distinct language models, reveals that the 1-hop expansion strategy consistently outperforms the baseline approach in subjective metrics, such as comprehensiveness and diversity, and frequently in qualitative metrics, like answer and contextual relevance.

**Keywords:** *GraphRAG*, *RAG*, *LLM*, *FHIR*, Graphs, Healthcare

# Sumário

<b>1</b>	<b>Introdução</b>	<b>1</b>
1.1	Justificativa . . . . .	2
1.2	Objetivos . . . . .	3
1.3	Hipótese . . . . .	4
1.4	Contribuições do trabalho . . . . .	4
1.5	Organização do Documento . . . . .	5
<b>2</b>	<b>Fundamentação Teórica</b>	<b>6</b>
2.1	Modelos de Linguagem de Grande Escala . . . . .	6
2.2	Geração Aumentada por Recuperação . . . . .	8
2.3	Geração Aumentada por Recuperação Baseada em Grafos . . . . .	10
2.4	Expansão de Contexto por Travessia de Grafo: O Algoritmo <i>K-Hop</i> . . . . .	12
2.4.1	Formalização Matemática . . . . .	13
2.4.2	Distinção de <i>Multi-hop Question Answering</i> . . . . .	14
2.5	O Desafio dos Contextos Longos e a Degradação da Atenção em <i>GraphRAG</i> . . . . .	14
2.5.1	O Fenômeno “ <i>Lost in the Middle</i> ” (Perdido no Meio) . . . . .	14
2.5.2	Implicações para a Arquitetura <i>GraphRAG</i> . . . . .	15
2.6	Métricas de avaliação em <i>RAG</i> . . . . .	16
2.6.1	Avaliação do recuperação . . . . .	16
2.6.2	Avaliação da geração . . . . .	17
2.7	<i>Fast Healthcare Interoperability Resources</i> . . . . .	20
<b>3</b>	<b>Trabalhos Relacionados</b>	<b>22</b>
<b>4</b>	<b>Proposta</b>	<b>27</b>
4.1	Base de Dados . . . . .	28
4.2	Construção da base de conhecimentos em grafos . . . . .	31
4.2.1	Principais Etapas do Pré-processamento . . . . .	31
4.3	Sistema de diálogo . . . . .	35
4.3.1	Fluxo da arquitetura . . . . .	35

4.3.2	Recuperação de Informação . . . . .	36
4.3.3	Formatação do grafo recuperado . . . . .	37
4.3.4	Experimentos . . . . .	39
4.4	Avaliação . . . . .	40
4.4.1	Métricas qualitativas . . . . .	41
4.4.2	Métricas subjetivas . . . . .	42
4.4.3	Métricas objetivas . . . . .	44
<b>5</b>	<b>Resultados</b>	<b>46</b>
5.1	Construção da base de grafos . . . . .	46
5.2	Experimentos . . . . .	47
5.3	Resultados gerais . . . . .	48
5.4	Resultados específicos . . . . .	52
5.5	Discussão . . . . .	54
5.5.1	Desempenho das Estratégias . . . . .	55
5.5.2	O Impacto do Perfil Clínico na Recuperação . . . . .	55
<b>6</b>	<b>Conclusões</b>	<b>57</b>
6.1	Contribuições do Trabalho . . . . .	58
6.2	Limitações do Estudo . . . . .	58
6.3	Trabalhos Futuros . . . . .	59
	<b>Referências</b>	<b>60</b>
	<b>Apêndice</b>	<b>64</b>
	<b>A Instrução de sistema</b>	<b>65</b>
	<b>B Instrução de Pre-processamento</b>	<b>66</b>
	<b>C Instrução de Avaliação</b>	<b>67</b>
	<b>D Principais Parâmetros</b>	<b>69</b>
	<b>E Exemplo de arquivo FHIR</b>	<b>71</b>
	<b>F Resultados por modelo</b>	<b>79</b>

# Lista de Figuras

2.1	Resumo das etapas de uma <i>RAG</i> [1]. . . . .	9
2.2	Fases da <i>GraphRAG</i> , desde a indexação de dados brutos até a geração de respostas aumentadas pelo grafo de conhecimento [2]. . . . .	10
4.1	Processo de geração da base de conhecimentos em grafos. . . . .	31
4.2	Exemplo dos relacionamos de um nó em específico. . . . .	34
4.3	Exemplo do nó persistido no banco de dados, onde na esquerda é possível ver as propriedades. . . . .	34
4.4	Arquitetura geral do sistema de diálogo proposto. À esquerda, estão os componentes de alto nível da arquitetura. . . . .	35
4.5	Representação gráfica das etapas da Busca Lexical. . . . .	36
4.6	Representação gráfica das etapas da Busca por Similaridade. . . . .	37
5.1	Representação gráfica da paciente <i>Beatris270 Bogan287</i> da Tabela 4.1. . . . .	47
5.2	Mapa de calor das médias de relevância da resposta (ARS) por paciente e por pergunta. . . . .	52
5.3	Mapa de calor das médias de relevância do contexto (CRS) por paciente e pergunta. . . . .	53
5.4	Mapa de calor das médias de Fidelidade $HAL^{-1}$ por paciente por pergunta. . . . .	54

# Lista de Tabelas

4.1	Resumo dos dados dos cinco pacientes selecionados do trabalho de Schmidmayer et al. (2024) [3]	29
4.2	Contagem dos Principais Tipos de Recursos e Total de Recursos <i>FHIR</i> por Paciente	30
4.3	Relação entre a quantidade de <i>tokens</i> e caracteres no arquivo de cada paciente.	30
4.4	Comparação entre recurso <i>FHIR</i> bruto e pré-processado. Na esquerda o recurso <i>FHIR</i> bruto, na direita o recurso <i>FHIR</i> pré processado. A instrução utilizada nessa etapa pode ser encontrada no Apêndice B	32
4.5	Comparações entre Formatos de Representação de Grafo no trabalho Guo et al. (2023) [4]. Na esquerda o formato de representação, na direita o exemplo correspondente.	38
4.6	Formato de <i>Adjacency List</i> utilizado no presente trabalho.	38
4.7	Perguntas avaliadas para cada paciente no presente trabalho.	40
4.8	Resultado para a pergunta Q1 sobre medicações utilizando o modelo <i>Gemini-1.5-pro</i> utilizando a busca lexical sem expansão de grafos para a paciente <i>Beatris270_Bogan287</i>	42
4.9	Resultados dos scores qualitativos para as entradas da Tabela 4.8	42
4.10	Entradas para a avaliação subjetiva, respostas L0 e L1 utilizando o modelo <i>Gemini-1.5-pro</i> para a paciente <i>Beatris270_Bogan287</i> após a consulta Q1	43
4.11	Avaliação subjetiva dada as entradas presentes na Tabela 4.10	44
4.12	Métricas de Tempo e <i>Tokens</i> por Tipo de Busca	45
5.1	Resultado da redução de <i>Tokens</i> por paciente após o pré processamento	48
5.2	Comparação de todas as estratégias por questão na escolha proporcional.	49
5.3	Comparação de todas as estratégias por paciente na escolha proporcional.	50
5.4	Comparação de estratégias de busca por pergunta	50
5.5	Comparação de estratégias de busca por Paciente	51
5.6	Estatísticas descritivas de Tempo e Tokens por Estratégia	51
A.1	Instrução de sistema para todos os experimentos	65

B.1	Instrução de pré-processamento do recurso FHIR em formato bruto para o formato de texto . . . . .	66
C.1	Instrução de avaliação utilizando LLM como julgador . . . . .	68
F.1	Comparação de estratégias de busca por modelo na escolha proporcional. . .	79
F.2	Comparação de métricas ARS, CRS e HALS por modelo e estratégia de busca . . . . .	80

# Capítulo 1

## Introdução

A saúde digital tem ganhado importância global crescente, com iniciativas que visam democratizar o acesso à informação em saúde. Programas como o *Healthy People 2030* [5], a introdução de APIs baseadas no padrão *Fast Healthcare Interoperability Resources* (FHIR) [6] e a promulgação do *21st Century Cures Act* nos Estados Unidos [7] destacam a necessidade crítica de eliminar disparidades em saúde, promover a equidade e melhorar a usabilidade e compreensão dos dados de saúde por todos os pacientes. A literacia em saúde, conforme definida pela *Health Resources & Services Administration* (HRSA), abrange não apenas o acesso à informação, mas também a capacidade de encontrar, compreender e utilizar de forma eficaz esses dados [8].

Contudo, a complexidade dos formatos técnicos e o uso de jargão médico especializado limitam significativamente a acessibilidade da informação em saúde. Isto é particularmente verdadeiro para pacientes com baixa literacia em saúde ou barreiras linguísticas [8]. Conseqüentemente, iniciativas destinadas a melhorar o acesso a Prontuários Eletrônicos de Pacientes e dados de saúde frequentemente não atingem o seu pleno potencial para otimizar os resultados clínicos dos pacientes. Neste contexto, os Modelos de Linguagem de Grande Escala (*LLMs*, do inglês, *Large Language Models*) [9] emergem como ferramentas transformadoras para superar estas barreiras de acessibilidade. Estudos recentes indicam que os *LLMs* podem interpretar dados médicos e gerar explicações em linguagem natural, tornando a informação de saúde mais acessível e compreensível [10, 11, 3, 12, 13, 14, 15, 16, 17]. Esta capacidade é especialmente relevante para reduzir disparidades em saúde e aumentar o impacto social. Apesar deste potencial, as abordagens tradicionais de Geração Aumentada por Recuperação (*RAG*, do inglês, *Retrieval-Augmented Generation*) enfrentam desafios ao lidar com as inter-relações complexas em dados clínicos.

Geração Aumentada por Recuperação Baseada em Grafo (*GraphRAG*, do inglês, *Graph-based Retrieval-Augmented Generation*) oferece um avanço significativo em relação a *RAG*

tradicional ao integrar *KG* e técnicas baseadas em grafos, como indexação, recuperação e geração [2]. Diferentemente da *RAG* tradicional, que frequentemente apresenta dificuldades com a natureza complexa e interconectada dos dados de saúde, como os recursos *FHIR* [10], a *GraphRAG* destaca-se por capturar relações e dependências. Isso facilita uma compreensão mais holística da informação do paciente, reduzindo a redundância e aumentando a relevância e a riqueza contextual das respostas geradas. Especificamente na área da saúde, abordagens como *MedGraphRAG* demonstram a capacidade da *GraphRAG* de ancorar as respostas dos *LLMs* em conhecimento médico externo, melhorando a confiabilidade e fornecendo informações de saúde baseadas em evidências [16].

Neste trabalho, empregamos um mecanismo de travessia de grafos fundamental: a expansão por saltos (*k-hop expansion*) para superar as limitações da busca lexical e por similaridade em prontuários eletrônicos no formato *FHIR*. Esta abordagem é crucial para capturar informações interconectadas, a fim de fornecer respostas mais abrangentes, detalhadas e diversificadas em comparação com métodos tradicionais. Especificamente, propomos e avaliamos uma estratégia de expansão de grafo de vizinhos de 1 salto (*1-hop neighbors*) sobre referências de busca lexical e busca por similaridade para aprimorar a recuperação de informações semanticamente relacionadas de *KGs* de prontuários eletrônicos. Para este fim, comparamos as versões propostas com as referências utilizando diversas métricas qualitativas e subjetivas, visando identificar os pontos fortes e fracos de cada abordagem.

Esta análise comparativa é conduzida utilizando dados *FHIR* sintéticos representando cinco pacientes distintos e um conjunto direcionado de oito perguntas (três de um artigo de referência [3] e cinco novas perguntas propostas) para uma avaliação mais abrangente, examinada em cinco *LLMs* de última geração. Propomos também uma avaliação ampla que incorpora métricas subjetivas e qualitativas, juntamente com a análise de latência e custo de *tokens*, para compreender a viabilidade da *GraphRAG* como sistema de diálogo. Desta forma, o objetivo é compreender as nuances de cada configuração e fornecer percepções de cada abordagem.

## 1.1 Justificativa

A progressiva digitalização na área da saúde tem resultado em um volume exponencial de dados clínicos. Embora estes dados detenham um potencial transformador, sua inerente complexidade e a utilização de formatos técnicos como *JSON* e *XML*, juntamente com terminologias médicas especializadas, impõem barreiras significativas à sua plena utilização tanto por pacientes quanto por profissionais de saúde [8]. Iniciativas globais, como o programa *Healthy People 2030* [5] e a padronização de *APIs* via *HL7 FHIR* [6],

buscam atenuar essas dificuldades, visando à democratização do acesso à informação e à promoção da literacia em saúde. No entanto, a efetividade dessas ações é frequentemente limitada pela dificuldade de pacientes, especialmente aqueles com baixo letramento médico ou barreiras linguísticas, em compreender e interagir com tais dados.

Nesse ínterim, *LLMs* surgiram como uma tecnologia promissora, capaz de interpretar dados complexos e traduzi-los para linguagem natural, tornando a informação médica mais acessível [10, 13]. Contudo, mesmo com o auxílio de técnicas como a *RAG*, persistem desafios notáveis, principalmente ao lidar com a natureza intrinsecamente interconectada dos dados em Prontuários Eletrônicos de Pacientes estruturados no padrão *FHIR*. Abordagens *RAG* convencionais frequentemente não exploram adequadamente essas ricas inter-relações, resultando em respostas que podem carecer de profundidade ou contexto completo [11, 3]. Esta limitação evidencia uma lacuna crítica no desenvolvimento de sistemas de diálogo clínico que possam verdadeiramente alavancar a riqueza semântica e estrutural dos dados de saúde.

A abordagem *GraphRAG*, ao integrar explicitamente a estrutura de *KG* ao processo da *RAG*, apresenta-se como uma solução inovadora e direcionada para superar tais desafios [2, 16]. Ao modelar e explorar as conexões entre diferentes recursos *FHIR*, a *GraphRAG* possui o potencial de fornecer respostas consideravelmente mais completas, precisas e contextualmente relevantes. Acredita-se que a investigação e validação da *GraphRAG* em dados *FHIR* podem não apenas aprimorar a qualidade da interação paciente-informação, mas também otimizar o fluxo de trabalho de profissionais de saúde.

Adicionalmente, a análise de aspectos práticos como latência e custo de *tokens* é fundamental para assegurar a viabilidade e escalabilidade dessas soluções em ambientes produtivos, contribuindo para um avanço tangível na forma como as informações de saúde são acessadas, compreendidas e utilizadas em benefício da equidade e eficácia no cuidado ao paciente.

## 1.2 Objetivos

Este trabalho tem como objetivo principal investigar e validar a eficácia da abordagem *GraphRAG* na interpretação de dados de Prontuários Eletrônicos de Pacientes no formato *FHIR*. Busca-se, fundamentalmente, determinar se a utilização de *GraphRAG* resulta em melhorias significativas na qualidade e relevância das informações recuperadas e geradas, conforme avaliado por métricas qualitativas e subjetivas, em comparação com técnicas de *RAG* tradicionais que não exploram explicitamente as inter-relações em grafos.

Adicionalmente, pretende-se analisar a robustez da abordagem *GraphRAG* no que tange à mitigação de alucinações e compreender as nuances de sua aplicação em um con-

texto de dados clínicos complexos e interconectados, contribuindo para um entendimento detalhado de seu potencial e limitações na literatura.

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Analisar e comparar o desempenho de abordagens de busca lexical e por similaridade, sem expansão de grafos, com a abordagem *GraphRAG* (utilizando expansão de 1 salto). Esta análise foca na capacidade de cada método em fornecer respostas diretas versus detalhadas para consultas que exigem a compreensão de informações inter-relacionadas, avaliando também o consumo de *tokens* e a suscetibilidade à degradação de contexto.
- Avaliar o impacto da incorporação de grafos na diversidade e abrangência das respostas geradas por sistemas de diálogo clínico. Investigar como a recuperação de informações baseada em grafos afeta a quantidade de *tokens* processados e as implicações para a implementação em larga escala, incluindo a potencial degradação da qualidade em interações de contexto longo.

### 1.3 Hipótese

Como hipótese principal, este trabalho postula que a implementação da abordagem *GraphRAG* para a interpretação de dados de Prontuários Eletrônicos de Pacientes em formato *FHIR* resultará em uma melhoria nas métricas subjetivas e qualitativas de avaliação, quando comparada às técnicas de *RAG* tradicionais, como verificado na literatura [18].

Adicionalmente, propõem-se as seguintes hipóteses secundárias para validação:

- Abordagens de busca lexical e por similaridade, quando aplicadas sem expansão de grafos, fornecerão respostas mais diretas e potencialmente menos detalhadas para consultas que envolvem informações inter-relacionadas.

### 1.4 Contribuições do trabalho

Como resultado direto do desenvolvimento e validação da arquitetura proposta, destaca-se a seguinte produção bibliográfica de nível internacional:

Baptista, João V. R. e Luís P. F. Garcia: *Graphrag on electronic health record: A knowledge graph-enhanced rag approach for healthcare information access*. Em Freitas, Rosiane de e Diego Furtado (editores): *Intelligent Systems*, páginas 615–630, Cham, 2026. Springer Nature Switzerland, ISBN 978-3-032-15987-8. [19].

## 1.5 Organização do Documento

Este trabalho está estruturado em seis capítulos, organizados da seguinte forma:

O Capítulo 1 introduz a pesquisa, estabelecendo a contextualização, os objetivos e as hipóteses norteadoras do estudo. O Capítulo 2 dedica-se à Revisão Bibliográfica, fundamentando teoricamente os conceitos de *LLMs*, *RAG*, *GraphRAG* e o padrão FHIR. No Capítulo 3, são discutidos os Trabalhos Relacionados, posicionando esta investigação frente ao estado da arte.

A metodologia e o desenvolvimento são detalhados no Capítulo 4, que abrange a arquitetura da solução, o processamento dos dados e o protocolo experimental. O Capítulo 5 expõe os Resultados e Discussões, apresentando as avaliações quantitativas e qualitativas que validam a proposta. Por fim, o Capítulo 6 encerra o trabalho com as considerações finais, sintetizando as contribuições alcançadas e sugerindo caminhos para trabalhos futuros.

# Capítulo 2

## Fundamentação Teórica

A seção a seguir dedica-se à revisão bibliográfica dos conceitos fundamentais que embasam este trabalho, com foco nas abordagens de *RAG* e suas extensões, como a *GraphRAG*. O objetivo é apresentar as técnicas e modelos que têm sido desenvolvidos para aprimorar a capacidade dos *LLMs* em tarefas que demandam conhecimento específico e atualizado, especialmente em domínios complexos como a saúde. A revisão está organizada de forma a explorar as etapas, limitações e avanços dessas abordagens, destacando sua relevância para aplicações práticas.

Inicialmente, são discutidos os fundamentos do *LLMs* na Seção 2.1, a base técnica do trabalho, em seguida, na Seção 2.2 apresenta-se a *RAG*, uma técnica que combina a recuperação de informações de bases de conhecimento externas com a geração de respostas por *LLMs*, mitigando problemas como alucinações e imprecisões. Em seguida, na Seção 2.3, são apresentadas as etapas da *GraphRAG*, uma evolução da *RAG* que utiliza estruturas de grafos para capturar relações complexas entre dados, oferecendo uma abordagem robusta para contextos especializados. A Seção 2.4 está apresentada os fundamentos do algoritmo de travessia utilizado no presente trabalho. Na Seção 2.5, é feita uma apresentação do problema de contextos longos que afetam a qualidade das respostas geradas pelos modelos. Assim como, são discutidas as métricas de avaliação e estratégias de avaliação utilizadas para mensurar a eficácia dessas técnicas na Seção 2.6. Na Seção 2.7, uma breve apresentação do modelo *FHIR* que é o padrão de dados de saúde utilizado no presente trabalho.

### 2.1 Modelos de Linguagem de Grande Escala

*LLMs* representam a vanguarda do processamento de linguagem natural, sendo o alicerce sobre o qual técnicas como *RAG* e *GraphRAG* são construídas. Um *LLM* é um modelo de aprendizado profundo treinado em vastos volumes de dados textuais, projetado para

compreender, gerar e interagir com a linguagem humana de maneira fluente e coerente. A escala desses modelos, frequentemente com bilhões de parâmetros, permite-lhes capturar padrões, nuances, fatos e estilos de raciocínio complexos presentes nos dados de treinamento [20].

O avanço fundamental que possibilitou a era dos *LLMs* foi a introdução da arquitetura *Transformer*, proposta por Vaswani et al. (2017) [9]. Diferentemente das arquiteturas recorrentes (RNNs) e convolucionais (CNNs) que a precederam, a arquitetura *Transformer* utiliza mecanismos de auto-atenção (*self-attention*) para processar sequências inteiras de texto de uma só vez. Esse mecanismo permite ao modelo ponderar a importância de diferentes palavras na sequência de entrada, independentemente de sua posição, capturando dependências de longo alcance e contextos complexos de forma muito mais eficaz.

A metodologia de treinamento predominante para *LLMs* segue um paradigma de duas fases: pré-treinamento e ajuste fino (*fine-tuning*). Durante o pré-treinamento, o modelo aprende a prever a próxima palavra em uma frase ou a preencher partes mascaradas do texto a partir de um corpus massivo e não rotulado. Essa fase dota o modelo de um conhecimento geral sobre o mundo e uma profunda competência linguística. Posteriormente, na fase de ajuste fino, o modelo pré-treinado é adaptado para tarefas específicas (como responder perguntas, traduzir ou sumarizar) ou para se alinhar com o comportamento desejado, utilizando conjuntos de dados menores e mais específicos, frequentemente com o auxílio de técnicas como o Aprendizado por Reforço com *Feedback* Humano (RLHF). Modelos proeminentes, como os da família GPT (*Generative Pre-trained Transformer*) [21], exemplificam o sucesso dessa abordagem.

Apesar de suas capacidades impressionantes, os *LLMs* possuem limitações intrínsecas que são críticas para aplicações em domínios especializados, como a saúde. Primeiramente, seu conhecimento é estático e limitado à data de corte de seus dados de treinamento, o que os torna incapazes de acessar informações em tempo real ou eventos recentes. Em segundo lugar, eles são propensos a “alucinações”, a geração de informações plausíveis, mas factualmente incorretas ou não fundamentadas nos dados de entrada [1]. Por fim, seu conhecimento, embora vasto, é generalista e pode carecer da profundidade e especificidade necessárias para domínios que exigem alta precisão, como o diagnóstico médico ou a interpretação de registros de saúde.

Essas limitações, em especial a propensão a alucinações e a falta de acesso ao conhecimento externo e atualizado, motivaram o desenvolvimento de novas arquiteturas que aumentam as capacidades dos *LLMs*. A *RAG* discutida na Seção 2.2 a seguir surge como uma solução direta para mitigar esses desafios, conectando os *LLMs* a bases de conhecimento externas e dinâmicas.

## 2.2 Geração Aumentada por Recuperação

Segundo Gao et al. (2023) [1], os modelos de linguagem têm alcançado avanços significativos, mas enfrentam limitações em tarefas específicas ou intensivas em conhecimento. Uma das principais falhas ocorre ao responder perguntas sobre tópicos além de seus dados de treinamento, o que pode resultar em “alucinações”. Para mitigar esse problema, a *RAG* complementa os *LLMs* com um mecanismo de recuperação de informações de bases de conhecimento externas.

O objetivo fundamental da *RAG* pode ser formalizado como a busca pela resposta ótima  $a^*$  que maximiza a probabilidade de uma resposta  $a$  ser correta, dada uma consulta  $q$  e uma base de conhecimento externa  $\mathcal{D}$ . Matematicamente, o problema é definido como:

$$a^* = \arg \max_{a \in \mathcal{A}} p(a|q, \mathcal{D}) \quad (2.1)$$

onde  $\mathcal{A}$  representa o espaço de todas as respostas possíveis. A abordagem *RAG* decompõe esse problema em duas etapas sequenciais: um recuperador e um gerador. Primeiro, o recuperador seleciona um subconjunto de documentos relevantes  $\mathcal{C}^* \subset \mathcal{D}$  e, em seguida, o gerador produz a resposta com base nesse contexto.

A Figura 2.1 ilustra o fluxo de uma *RAG* tradicional, que pode ser formalmente dividido em três etapas principais: Indexação, Recuperação e Geração.

- **Indexação:** Esta etapa prepara a base de conhecimento externa. Uma coleção de documentos brutos é processada, limpa e segmentada em blocos de texto menores,  $\mathcal{D} = \{c_1, c_2, \dots, c_N\}$ . Em seguida, um modelo de vetorial pré-treinado,  $E : \mathcal{T} \rightarrow R^d$ , mapeia cada bloco de texto  $c_i$  do espaço de texto  $\mathcal{T}$  para um vetor denso  $v_i$  em um espaço vetorial de dimensão  $d$ .

$$v_i = E(c_i) \quad (2.2)$$

O conjunto de todos os vetores,  $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ , é então armazenado e indexado em um banco de dados vetorial, que permite buscas de similaridade de forma eficiente.

- **Recuperação:** Dada uma consulta do usuário  $q$ , esta etapa visa encontrar os blocos de informação mais relevantes na base de conhecimento indexada. A consulta  $q$  é primeiro transformada em um vetor  $q_v$  usando o mesmo modelo de vetorial:  $q_v = E(q)$ . O sistema então busca os  $k$  vetores de blocos mais próximos de  $q_v$  no espaço vetorial. Formalmente, o conjunto de bloco recuperado,  $\mathcal{C}_k^*$ , é obtido por:

$$\mathcal{C}_k^* = \text{top-k}_{c_i \in \mathcal{D}} \text{Sim}(E(q), E(c_i)) \quad (2.3)$$

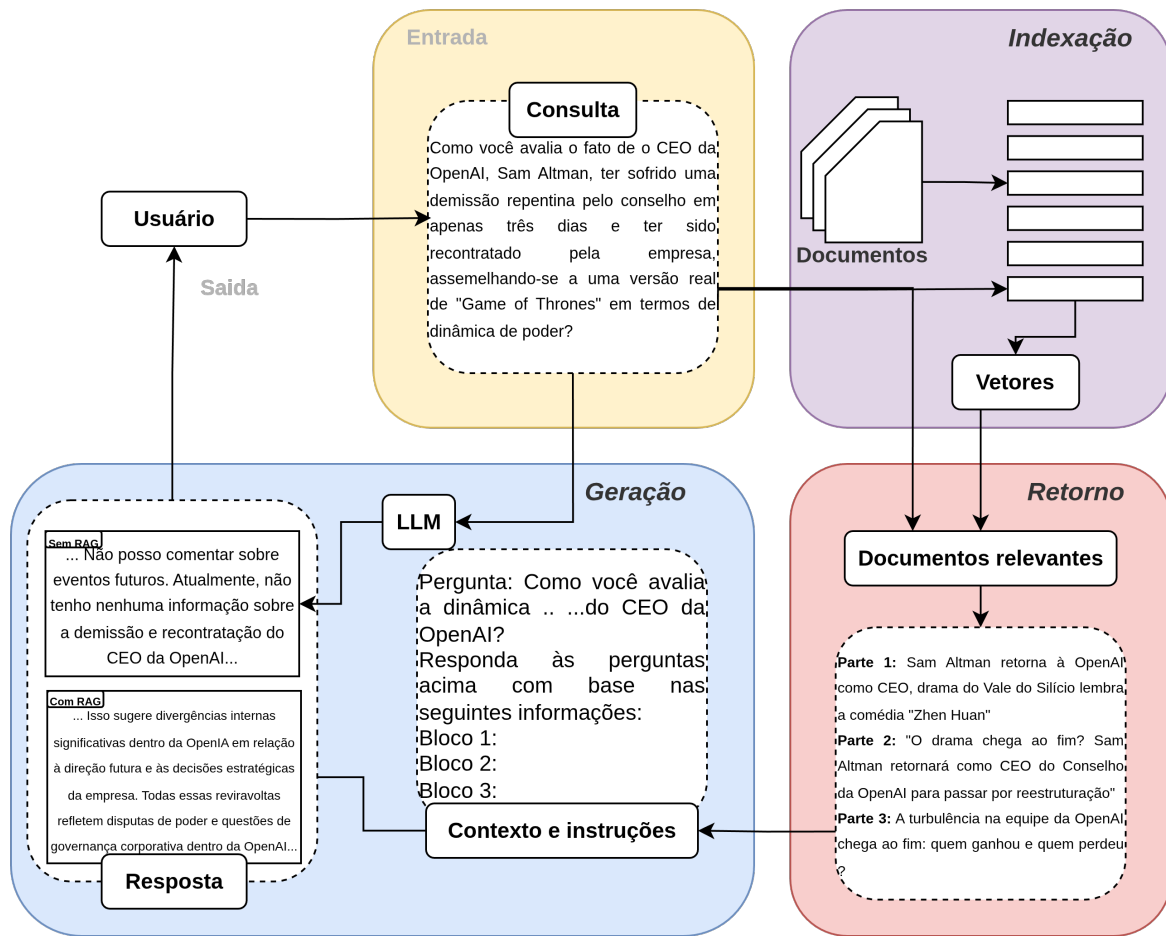


Figura 2.1: Resumo das etapas de uma *RAG* [1].

onde  $\text{Sim}(u, v)$  é uma função de similaridade, como a similaridade de cossenos  $(\frac{u \cdot v}{\|u\| \|v\|})$ , o produto interno  $(u \cdot v)$  ou a distância Euclidiana negativa  $(-\|u - v\|)$ . O conjunto  $\mathcal{C}_k^*$  forma o contexto aumentado que será usado na etapa de geração.

- **Geração:** Na etapa final, o modelo de linguagem, parametrizado por  $\phi$ , utiliza o contexto recuperado  $\mathcal{C}_k^*$  para gerar uma resposta final coerente e contextualizada. A consulta original  $q$  e o contexto  $\mathcal{C}_k^*$  são combinados em um consulta aumentado,  $P_{aug} = F(q, \mathcal{C}_k^*)$ , onde  $F$  é uma função de formatação. A resposta final  $a^*$  é gerada ao modelar a distribuição de probabilidade da próxima palavra, condicionada a essa consulta:

$$a^* = \arg \max_{a \in \mathcal{A}} p_\phi(a | F(q, \mathcal{C}_k^*)) \quad (2.4)$$

Apesar de sua simplicidade conceitual, esse processo apresenta desafios significativos [1]. Problemas na precisão da recuperação podem levar à seleção de informações irrelevantes, à omissão de dados cruciais ou ao problema de “contexto perdido no meio” em

janelas de contexto muito longas [22].

Durante a geração, persistem os riscos de alucinações, irrelevâncias e vieses, que afetam a qualidade das respostas. Além disso, a integração das informações recuperadas pode gerar redundância ou incoerência estilística. Essas limitações destacam a necessidade de estratégias avançadas para melhorar a precisão e a relevância na *RAG* tradicional, especialmente em tarefas críticas como o processamento de dados em aplicações de saúde.

## 2.3 Geração Aumentada por Recuperação Baseada em Grafos

*GraphRAG* representa uma evolução da abordagem de *RAG*, ao incorporar estruturas de grafos para aprimorar a organização, recuperação e geração de informações [2]. A transição para a *GraphRAG* fundamenta-se na capacidade dos grafos de capturar eficientemente as relações e dependências intrínsecas presentes em conjuntos de dados complexos [23].

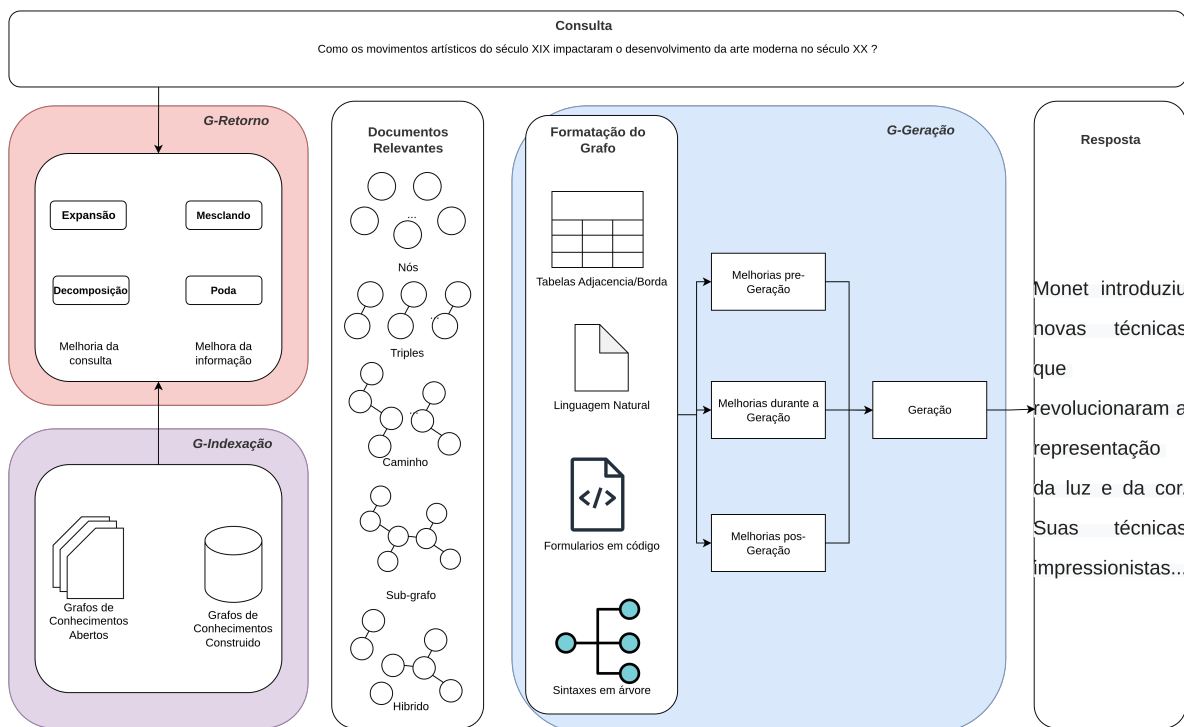


Figura 2.2: Fases da *GraphRAG*, desde a indexação de dados brutos até a geração de respostas aumentadas pelo grafo de conhecimento [2].

O objetivo fundamental da *GraphRAG* pode ser formalizado como a busca pela resposta ótima  $a^*$  que maximiza a probabilidade condicional, dada uma consulta  $q$  e uma

base de conhecimento em grafo  $G$ . Matematicamente, o problema é definido como:

$$a^* = \arg \max_{a \in A} p(a|q, G) \quad (2.5)$$

Esta decomposição estrutura a arquitetura da *GraphRAG* em três componentes sequenciais e interdependentes, conforme ilustrado na Figura 2.2: Indexação baseada em Grafos, Retorno baseado em Grafos e Geração baseada em Grafos.

- **Indexação baseada em Grafos:** Esta etapa inicial consiste na construção de uma base de KG, formalmente representada como  $G = (V, E)$ , onde  $V$  é o conjunto de nós (entidades) e  $E$  é o conjunto de arestas (relações). Os dados brutos são processados e transformados nesta estrutura, e a criação de índices eficientes sobre este grafo é fundamental para permitir a organização e a subsequente recuperação de informações de maneira que preserve suas complexas interconexões.
- **Retorno baseado em Grafos:** Nesta fase, o sistema utiliza a estrutura do grafo para aprimorar a precisão e a relevância da recuperação de dados. Formalmente, o objetivo é encontrar um subgrafo ótimo  $G^*$  a partir de uma consulta  $q$ , o que pode ser expresso como:

$$G^* = \text{G-Retriever}(q, G) = \arg \max_{G' \subseteq R(G)} \text{Sim}(q, G') \quad (2.6)$$

Aqui,  $\text{Sim}(q, G')$  é uma função que mede a similaridade semântica entre a consulta e um subgrafo candidato  $G'$ , e  $R(G)$  é uma função que restringe o espaço de busca para garantir a eficiência. Em vez de uma simples busca por palavras-chave, são empregados algoritmos de travessia e busca em grafos para implementar essa busca, como buscas de vizinhança, busca em largura (*Breadth-First Search*, BFS) e busca em profundidade (*Depth-First Search*, DFS) [2]. A recuperação pode ser configurada com diferentes níveis de granularidade, como nós individuais, triplas (sujeito-predicado-objeto) ou subgrafos inteiros.

- **Geração baseada em Grafos:** A etapa final aproveita o subgrafo recuperado  $G^*$  para enriquecer o processo de geração de texto. As informações estruturadas do grafo são fornecidas como entrada para um *LLM*. A geração da resposta final  $a^*$  é modelada como:

$$a^* = \text{G-Generator}(q, G^*) = \arg \max_{a \in A} p_\phi(a|F(q, G^*)) \quad (2.7)$$

A função  $F(q, G^*)$  é responsável por converter os dados do grafo em um formato que o *LLM* possa processar de forma eficaz. Em arquiteturas avançadas, Redes Neurais

em Grafos (*Graph Neural Networks*, GNNs) podem ser usadas para implementar  $F$ , pré-processando ou codificando as informações do grafo em um formato vetorial que o *LLM* interpreta de forma mais eficaz.

A flexibilidade da *GraphRAG* permite a incorporação de técnicas de otimização antes, durante e após a geração, visando aprimorar a qualidade dos resultados finais. Em um estudo relevante, Guo et al. (2023) realizaram uma análise comparativa metódica sobre a capacidade interpretativa dos *LLMs* ao processar dados estruturados em grafos [4]. O trabalho investiga como diferentes estruturas de grafos, formatos de entrada e estratégias de *prompting* influenciam a precisão dos resultados gerados. Ao distinguir entre tarefas de compreensão estrutural e semântica, a pesquisa explora o potencial dos *LLMs* para além dos dados tabulares convencionais, evidenciando a complexidade inerente à interpretação de informações representadas por nós e arestas [4].

Enquanto a *RAG* tradicional foca na recuperação de informações baseada em textos e na subsequente geração de conteúdo, a *GraphRAG* expande esta abordagem ao incorporar grafos como núcleo central do processamento das informações. Segundo Peng et al. (2024) e Edge et al. (2024), uma das principais limitações da *RAG* é a sua incapacidade de capturar inter-relações complexas nos dados, algo que os grafos conseguem fazer com eficácia [2, 18]. Na *GraphRAG*, a utilização de grafos permite um entendimento mais profundo das dependências e das conexões entre diferentes entidades, reduzindo informações redundantes e englobando informações globais de maneira mais eficiente do que na configuração *RAG*. Este diferencial torna a *GraphRAG* especialmente relevante para contextos de saúde, onde as interações entre dados clínicos podem ser não lineares e altamente complexas [2, 18].

## 2.4 Expansão de Contexto por Travessia de Grafo: O Algoritmo *K-Hop*

Na etapa de recuperação de uma arquitetura *GraphRAG*, a obtenção de um contexto rico e interconectado é fundamental para a qualidade da resposta final. Frequentemente, a recuperação inicial, baseada em similaridade vetorial, retorna apenas um conjunto limitado de nós de entrada ( $V_0$ ), que podem ser insuficientes para representar a totalidade da informação relevante. Para superar essa limitação, são empregados algoritmos de travessia de grafo que expandem o contexto a partir desses nós iniciais. Um dos algoritmos mais fundamentais e amplamente utilizados para essa finalidade é a expansão da vizinhança de *k-hop* [2].

O algoritmo opera sobre um grafo de conhecimento  $G = (V, E)$  para coletar um subgrafo de contexto,  $G_{ctx}$ , a partir de um conjunto inicial de nós  $V_0 \subseteq V$ . A intuição é que a informação mais relevante para uma consulta tende a estar estruturalmente próxima dos nós diretamente identificados. A expansão k-hop, portanto, coleta todos os nós que estão a uma distância de até  $k$  arestas (ou “saltos”) dos nós em  $V_0$ .

### 2.4.1 Formalização Matemática

Dado um grafo de conhecimento  $G = (V, E)$ , onde  $V$  é o conjunto de nós (entidades) e  $E$  é o conjunto de arestas (relações), e um conjunto de nós iniciais  $V_0 \subseteq V$ , o algoritmo de expansão k-hop constrói iterativamente um conjunto de nós expandidos.

Seja  $N(v)$  a função que retorna a vizinhança imediata (1-hop) de um nó  $v \in V$ :

$$N(v) = \{u \in V \mid (v, u) \in E\} \quad (2.8)$$

A vizinhança de 1-hop de um conjunto de nós  $S \subseteq V$  é a união das vizinhanças de cada nó em  $S$ :

$$N(S) = \bigcup_{v \in S} N(v) \quad (2.9)$$

O conjunto de nós da vizinhança de k-saltos,  $V_k$ , pode ser definido recursivamente:

- **Base (k=0):** O conjunto inicial de nós é  $V_0$ .
- **Passo Recursivo (k>0):** A vizinhança de k-saltos é obtida expandindo a vizinhança de  $(k - 1)$ -saltos, incluindo todos os nós adjacentes.

$$V_k = V_{k-1} \cup N(V_{k-1}) \quad (2.10)$$

Após  $k$  iterações, o conjunto final de nós  $V_k$  contém todos os nós que estão a uma distância de caminho  $\leq k$  de pelo menos um nó em  $V_0$ . O subgrafo de contexto final,  $G_k = (V_k, E_k)$ , é então induzido pelo conjunto de nós  $V_k$ , onde  $E_k$  é o conjunto de todas as arestas de  $G$  que conectam dois nós em  $V_k$ :

$$E_k = \{(u, v) \in E \mid u \in V_k \wedge v \in V_k\} \quad (2.11)$$

Este subgrafo  $G_k$  é então fornecido ao *LLM* como o contexto aumentado para a geração da resposta. A escolha do hiperparâmetro  $k$  representa um balanço crítico: valores pequenos podem levar a um contexto insuficiente, enquanto valores grandes podem introduzir ruído e aumentar o custo computacional.

## 2.4.2 Distinção de *Multi-hop Question Answering*

É crucial distinguir a expansão  $k$ -hop em grafos do conceito de *Multi-hop Question Answering (QA)*, uma tarefa consolidada em Processamento de Linguagem Natural (NLP) [24]. Embora ambos os termos envolvam “saltos” (hops), eles se referem a processos fundamentalmente distintos:

***Multi-hop QA*** refere-se a uma tarefa de raciocínio, na qual um sistema deve sintetizar informações de múltiplos documentos ou passagens de texto para responder a uma pergunta complexa. Os “saltos” aqui são passos lógicos ou inferenciais que conectam peças de evidência textual dispersas. A estrutura subjacente é implícita e deve ser descoberta pelo modelo durante o processo de raciocínio. O objetivo é a resposta final.

***K-hop Expansion*** refere-se a um algoritmo de travessia estrutural em uma base de conhecimento explícita, como um grafo. Os “saltos” são travessias físicas ao longo de arestas pré-definidas. O objetivo não é responder à pergunta diretamente, mas sim coletar um subgrafo de contexto relevante. O raciocínio complexo é delegado ao *LLM*, que recebe este subgrafo como entrada.

Em suma, enquanto o *Multi-hop QA* é sobre o processo de inferência para *encontrar* a resposta, a expansão  $k$ -hop na *GraphRAG* é sobre o processo de coleta de dados para *contextualizar* a geração da resposta.

## 2.5 O Desafio dos Contextos Longos e a Degradação da Atenção em *GraphRAG*

Um dos principais objetivos da *GraphRAG* é superar a limitação do *RAG* tradicional, que trata a informação como fragmentos isolados. Ao empregar algoritmos de travessia como a expansão  $k$ -hop, a *GraphRAG* constrói um subgrafo de contexto potencialmente vasto e rico em interconexões. No entanto, essa capacidade introduz um dos maiores desafios da geração de linguagem moderna: o problema dos contextos longos (*long-context problem*) [1]. A premissa de que “mais contexto é sempre melhor” é falaciosa; a inclusão indiscriminada de informação pode degradar, em vez de aprimorar, a qualidade da resposta gerada.

### 2.5.1 O Fenômeno “*Lost in the Middle*” (Perdido no Meio)

Pesquisas recentes sobre o comportamento de *LLMs* com janelas de contexto extensas revelaram um padrão de atenção inconsistente. O fenômeno, popularmente conhecido

como “agulha no palheiro” (*needle in a haystack*) e formalmente analisado como “perdido no meio” (*lost in the middle*), demonstra que a capacidade de um *LLM* de identificar e utilizar uma informação específica depende fortemente de sua posição dentro do contexto [25].

Especificamente, os modelos tendem a dar maior atenção às informações localizadas no início e no fim da janela de contexto, enquanto a performance decai significativamente para informações localizadas na região central. Se um fato crucial ou nó relevante for inserido no meio de um longo contexto, a probabilidade de o modelo “esquecê-lo” ou ignorá-lo aumenta drasticamente. Este comportamento pode ser visualizado como uma curva de performance em formato de “U”, onde a precisão é alta nas extremidades e baixa no centro.

### 2.5.2 Implicações para a Arquitetura *GraphRAG*

A natureza da recuperação em *GraphRAG* exacerba diretamente o risco de degradação contextual. O processo de expansão do grafo, ao coletar nós e suas respectivas descrições textuais, gera um volume de dados que deve ser linearizado — ou seja, convertido em uma sequência de texto — para ser inserido no contexto do *LLM*. Este processo de linearização cria vários pontos de falha:

- **Ordenação Arbitrária:** A conversão de um subgrafo não-linear para uma sequência de texto linear pode posicionar nós de alta relevância, mas que foram descobertos em saltos intermediários (e.g., 2-hop), na região central do contexto. Com isso, o benefício de uma recuperação de grafo sofisticada é anulado pela incapacidade do *LLM* de “perceber” a informação crucial.
- **Trade-off entre Revocação e Ruído:** Aumentar o valor de  $k$  na expansão  $k$ -hop melhora a revocação (*recall*) na etapa de recuperação, capturando mais entidades potencialmente relevantes. Contudo, isso infla o contexto, aumentando a probabilidade de que informações críticas se percam no meio e introduzindo ruído que pode confundir o gerador.
- **Diluição da Relevância:** Um contexto excessivamente longo pode diluir a importância dos nós mais pertinentes. O *LLM* pode ter dificuldade em distinguir o sinal (informação-chave) do ruído (informação marginalmente relevante), resultando em respostas genéricas, incompletas ou que falham em sintetizar as conexões complexas capturadas pelo grafo.

## 2.6 Métricas de avaliação em *RAG*

Segundo o Gao et al. (2023), as avaliações de modelos *RAG* têm-se concentrado em seu desempenho em etapas específicas, utilizando métricas consolidadas adequadas para cada etapa. De acordo com os trabalhos de referencia, pode-se classificar em avaliações da busca, ou seja, se os mecanismos de buscas retornam informações pertinentes para o contexto da *LLM*. A etapa de geração, ou seja, a qualidade da resposta final fornece pelo modelo, dado as instruções e contexto fornecido.

### 2.6.1 Avaliação do recuperação

A qualidade da recuperação é um aspecto crucial para determinar a eficácia do contexto obtido na etapa de recuperação. Para medir o desempenho deste módulo nos modelos *RAG*, são utilizadas métricas padrão dos domínios de busca, sistemas de recomendação e recuperação de informação [26]. A formalização destas métricas pressupõe um conjunto de consultas de avaliação  $Q$ . Para cada consulta  $q \in Q$ , o sistema retorna uma lista ranqueada de  $K$  documentos,  $R_q = (d_1, d_2, \dots, d_K)$ , e existe um conjunto de documentos de referência (ground truth) considerados relevantes,  $D_q^+$ .

A seguir, são definidas formalmente as principais métricas: Hit Rate, MRR e NDCG.

**Hit Rate@K (Taxa de Acerto@K)** Mede a fração de consultas para as quais pelo menos um documento relevante foi recuperado entre os  $K$  primeiros resultados. É uma métrica binária de sucesso por consulta. Sua formulação é dada por:

$$\text{Hit Rate@K} = \frac{1}{|Q|} \sum_{q \in Q} I(R_q \cap D_q^+ \neq \emptyset) \quad (2.12)$$

onde  $I(\cdot)$  é a função indicadora, que retorna 1 se a condição for verdadeira (a intersecção não é vazia) e 0 caso contrário.

**Mean Reciprocal Rank (MRR)** Avalia a posição do primeiro documento relevante na lista ranqueada. O *Reciprocal Rank* para uma única consulta  $q$  é o inverso da posição (rank) do primeiro documento relevante. O MRR é a média desses valores sobre todas as consultas. Formalmente:

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}(q)} \quad (2.13)$$

onde  $\text{rank}(q) = \min\{i \mid d_i \in R_q \wedge d_i \in D_q^+\}$ . Se nenhum documento relevante for encontrado na lista,  $\frac{1}{\text{rank}(q)}$  é definido como 0.

**Normalized Discounted Cumulative Gain@K (NDCG@K)** É uma métrica que considera a relevância gradual dos documentos e penaliza a recuperação de itens relevantes em posições inferiores na lista. O cálculo envolve três passos [27]:

1. **Discounted Cumulative Gain (DCG)**: Soma os ganhos de relevância, descontados logaritmicamente pela posição  $i$ .

$$\text{DCG}_K(q) = \sum_{i=1}^K \frac{\text{rel}(d_i, q)}{\log_2(i + 1)} \quad (2.14)$$

onde  $\text{rel}(d_i, q)$  é o score de relevância (e.g., 0, 1, 2) do documento  $d_i$  para a consulta  $q$ .

2. **Ideal Discounted Cumulative Gain (IDCG)**: É o DCG da lista de resultados ideal, ou seja, com os documentos mais relevantes nas primeiras posições.

$$\text{IDCG}_K(q) = \sum_{i=1}^K \frac{\text{rel}^*(d_i, q)}{\log_2(i + 1)} \quad (2.15)$$

onde  $\text{rel}^*$  são os scores de relevância da lista perfeitamente ordenada.

3. **NDCG**: É a razão entre o DCG e o IDCG, normalizando o score para o intervalo  $[0, 1]$ . O valor final é a média sobre todas as consultas.

$$\text{NDCG@K} = \frac{1}{|Q|} \sum_{q \in Q} \frac{\text{DCG}_K(q)}{\text{IDCG}_K(q)} \quad (2.16)$$

A utilização conjunta dessas métricas oferece uma visão abrangente sobre a capacidade do sistema de encontrar documentos relevantes (Hit Rate), a rapidez com que os encontra (MRR) e a qualidade do ranqueamento desses documentos (NDCG).

## 2.6.2 Avaliação da geração

A avaliação da qualidade da resposta gerada é um pilar para mensurar a eficácia dos modelos RAG. As estratégias de avaliação podem ser categorizadas em duas abordagens principais: avaliações com rótulo de comparação, que dependem de respostas de referência (ground truth), e avaliações sem rótulo de comparação, que avaliam aspectos como a fidelidade ao contexto ou utilizam outros modelos de linguagem como juízes.

### Avaliação com Rótulo de Comparação

A avaliação com rótulo de comparação quantifica o desempenho de um modelo comparando a resposta gerada ( $a$ ) com uma ou mais respostas de referência pré-definidas ( $A^*$ ).

A escolha da métrica depende intrinsecamente da tarefa. Para resposta a perguntas, escores como Exact Match (EM) e F1-score são padrões. Tarefas de checagem de fatos, modeladas como classificação, utilizam a Acurácia. De forma mais ampla, para avaliar a sobreposição lexical e semântica, métricas como BLEU [28] e ROUGE [29] são comuns.

**Acurácia** Utilizada em tarefas de classificação (e.g., checagem de fatos binária), mede a proporção de predições corretas em relação ao total de amostras. Para um conjunto de  $N$  predições, é definida como:

$$\text{Acurácia} = \frac{\text{Número de Predições Corretas}}{\text{Número Total de Predições}} = \frac{1}{N} \sum_{i=1}^N I(a_i = a_i^*) \quad (2.17)$$

onde  $I(\cdot)$  é a função indicadora,  $a_i$  é a predição do modelo e  $a_i^*$  é o rótulo verdadeiro para a amostra  $i$ .

**Exact Match (EM)** Métrica binária que atribui uma pontuação de 1 se a resposta gerada  $a$  for idêntica, caractere por caractere, a uma das respostas de referência  $a^* \in A^*$ , e 0 caso contrário. A média é calculada sobre o conjunto de teste.

$$\text{EM}(a, A^*) = \max_{a^* \in A^*} I(a = a^*) \quad (2.18)$$

**F1-score** Mede a sobreposição de tokens entre a resposta gerada  $a$  e a referência  $a^*$ , tratando ambas como um *bag-of-tokens*. É a média harmônica entre a Precisão (P) e a Revocação (R):

$$\text{Precisão (P)} = \frac{|\text{Tokens}(a) \cap \text{Tokens}(a^*)|}{|\text{Tokens}(a)|} \quad (2.19)$$

$$\text{Revocação (R)} = \frac{|\text{Tokens}(a) \cap \text{Tokens}(a^*)|}{|\text{Tokens}(a^*)|} \quad (2.20)$$

$$\text{F1} = 2 \cdot \frac{P \cdot R}{P + R} \quad (2.21)$$

Quando múltiplas referências existem, calcula-se o F1 contra cada uma e retém-se o valor máximo.

**BLEU (Bilingual Evaluation Understudy)** Mede a precisão de n-gramas, comparando quantos n-gramas (sequências de  $n$  palavras) da resposta gerada aparecem nas referências. Inclui uma penalidade de brevidade (BP) para respostas muito curtas. A fórmula é:

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right) \quad (2.22)$$

onde  $p_n$  é a precisão dos n-gramas modificada e  $w_n$  são pesos (tipicamente uniformes,  $1/N$ ).

**ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** Avalia a qualidade de resumos e textos gerados com base na revocação de n-gramas. A variante mais comum, ROUGE-N, mede quantos n-gramas da(s) resposta(s) de referência aparecem na resposta gerada:

$$\text{ROUGE-N} = \frac{\sum_{s \in A^*} \sum_{\text{n-gram} \in s} \text{Count}_{\text{match}}(\text{n-gram})}{\sum_{s \in A^*} \sum_{\text{n-gram} \in s} \text{Count}(\text{n-gram})} \quad (2.23)$$

onde  $\text{Count}_{\text{match}}$  é o número de n-gramas na resposta gerada que também estão na referência.

### Avaliação sem Rótulo de Comparação

Quando respostas de referência não estão disponíveis ou são insuficientes para capturar qualidades como a fidelidade factual, recorre-se a métodos de avaliação sem rótulo. Estas abordagens avaliam a qualidade da geração com base na consistência interna entre a pergunta, o contexto recuperado e a resposta. As métricas focam em aspectos como a relevância do contexto, a fidelidade da resposta ao contexto e a robustez contra fatos contraditórios.

- **Context Relevance:** Avalia a precisão e a concisão do contexto recuperado,  $c(q)$ , garantindo que contenha apenas as informações essenciais para responder à pergunta. Um *LLM* pode ser usado para extrair um subconjunto de sentenças  $S_{ext}$  essenciais. A pontuação de relevância do contexto ( $CR$ ) é a razão entre o número de sentenças extraídas e o total de sentenças no contexto:

$$CR = \frac{|S_{ext}|}{|S_{c(q)}|} = \frac{\text{número de sentenças extraídas}}{\text{número total de sentenças em } c(q)} \quad (2.24)$$

- **Answer Faithfulness:** Garante que as respostas geradas sejam factualmente consistentes com o contexto recuperado. Para calcular a fidelidade, a resposta é decomposta em um conjunto de afirmações  $S$ . Cada afirmação é então verificada em relação ao contexto para determinar se é suportada. A pontuação de fidelidade ( $F$ ) é a razão entre o número de afirmações verificadas ( $|V|$ ) e o número total de afirmações ( $|S|$ ):

$$F = \frac{|V|}{|S|} \quad (2.25)$$

- **Answer Relevance:** Mede se a resposta gerada é pertinente à pergunta original. Para estimar esta relevância, um *LLM* gera  $n$  perguntas potenciais  $\{q_1, \dots, q_n\}$  a partir da resposta. A pontuação de relevância ( $AR$ ) é a média da similaridade de cossenos entre o embedding da pergunta original ( $q$ ) e os das perguntas geradas ( $q_i$ ):

$$AR = \frac{1}{n} \sum_{i=1}^n \text{sim}(q, q_i) \quad (2.26)$$

- **Hallucination Rate:** Determina a frequência com que o modelo gera informações que contradizem o contexto fornecido [30]. A métrica utiliza um *LLM* para identificar contradições entre a resposta e o contexto. O índice de alucinação é a razão entre o número de contextos contraditórios e o total de contextos analisados:

$$\text{Hallucination Rate} = \frac{\text{Número de Contextos Contraditórios}}{\text{Número Total de Contextos}} \quad (2.27)$$

### Avaliação por *LLMs* Julgadores (*LLM-as-a-Judge*)

Uma vertente proeminente da avaliação sem rótulo é o uso de *LLMs* como julgadores, que avaliam a qualidade das respostas de forma comparativa ou com base em critérios qualitativos. O trabalho de Zheng et al. (2023) introduziu o *MT-Bench*, uma plataforma de benchmark para avaliar *LLMs* em tarefas de julgamento, destacando suas capacidades e vieses intrínsecos [31].

Seguindo essa abordagem, trabalhos como os de Edge et al. (2024) e Guo et al. (2024) utilizaram *LLMs* para realizar comparações par a par entre respostas, avaliando-as segundo métricas qualitativas [18, 32] utilizando a instrução presente no Apêndice C:

- **Comprehensiveness:** Quantos detalhes a resposta fornece para cobrir todos os aspectos da pergunta.
- **Diversity:** Quão variada e rica é a resposta no fornecimento de diferentes perspectivas e informações.
- **Empowerment:** Até que ponto a resposta auxilia o leitor a compreender e formar julgamentos informados.
- **Directness:** Quão específica e claramente a resposta aborda a questão.

## 2.7 *Fast Healthcare Interoperability Resources*

O *FHIR* introduzido pelo trabalho de Bender et al. (2013), é um padrão desenvolvido pela HL7 para promover a troca de informações eletrônicas de saúde [6]. Desde sua introdução,

o *FHIR* tem se destacado por sua capacidade de integrar dados médicos de maneira eficiente, combinando funcionalidades dos padrões HL7 anteriores, como o HL7 v2 e o *HL7 Clinical Document Architecture* (CDA). Segundo Mandel et al. (2016), a principal meta do *FHIR* é facilitar a interoperabilidade entre diferentes sistemas de saúde, promovendo a troca eficaz de informações de saúde eletrônicas e, assim, melhorando o processo de tomada de decisão clínica [33]. Um exemplo pequeno desse padrão no formato em JSON pode ser visto no Apêndice E.

*FHIR* utiliza uma abordagem baseada em recursos, em que a informação médica é dividida em componentes modulares pequenos, denominados de recursos, como por exemplo; pacientes, observações e prescrições de medicamentos. Essa abordagem modular permite maior flexibilidade e personalização na maneira como os dados podem ser manipulados e compartilhados entre sistemas. *FHIR* também adota interfaces de programação de aplicativos (APIs) *RESTful*, que são bem integradas com JSON, XML e RDF, facilitando a comunicação rápida e escalável entre sistemas.

A comunidade *FHIR* definiu mais de 150 categorias de recursos diferentes [34], esses recursos são relacionados em outros recursos, formando uma rede interconectada de informações sobre o prontuário de um paciente. Entre os principais recursos, destacam-se:

- *Patient*: Contém dados demográficos, detalhes de contato e informações essenciais sobre os pacientes;
- *Observation*: Representa medições e achados clínicos, como pressão arterial e resultados de exames laboratoriais;
- *Condition*: Registra informações detalhadas sobre uma condição, problema, diagnóstico ou outro evento, situação ou problema que atingiu algum nível de preocupação;
- *AllergyIntolerance*: Avaliação clínica de alergia ou intolerância, propensão ou risco potencial para o um indivíduo;
- *MedicationRequest*: Este recurso é utilizado para rastrear medicamentos prescritos aos pacientes.

# Capítulo 3

## Trabalhos Relacionados

A aplicação de *RAG* na área da saúde tem sido amplamente explorada, com diversos estudos destacando seu potencial e suas limitações. Por exemplo, Ge et al. (2023) utilizaram *RAG* para desenvolver um aplicativo de chat especializado em doenças do fígado, reportando resultados promissores, mas observando limitações relacionadas aos conjuntos de dados empregados [12]. Veen et al. (2024) aplicaram *LLMs* na sumarização de dados médicos, alcançando níveis de desempenho comparáveis ou superiores aos de profissionais médicos, embora tenham reconhecido o risco persistente de alucinações [13].

Na mesma linha de *RAG*, Zaretsky et al. (2024) focaram na simplificação de resumos de alta hospitalar para pacientes, melhorando a legibilidade, mas encontrando desafios em manter a precisão e a completude [14]. Saba et al. (2024) e Elgedawy et al. (2024) exploraram ainda mais o uso de *RAG* em documentos médicos eletrônicos. Enquanto Saba et al. (2024) se concentraram na sumarização [15], Elgedawy et al. (2024) focaram na extração de informações, ambos utilizando metodologias de busca por similaridade para aprimorar a utilidade dos dados médicos [17].

Abordagens baseadas em grafos, como a *GraphRAG*, têm emergido como ferramentas poderosas para melhorar o desempenho de *LLMs* na saúde. Wu et al. (2024) introduziram o *MedGraphRAG*, um novo método baseado em grafos projetado para o domínio médico [16]. O *MedGraphRAG* aprimora as respostas de *LLMs* ao fundamentá-las em evidências e fornecer explicações de termos médicos oficiais. Essa abordagem constrói grafos de triplas que conectam dados de saúde a artigos médicos autorizados e dicionários especializados, garantindo confiabilidade e aplicabilidade.

Ao utilizar a estratégia *U-Retrieval*, o *MedGraphRAG* otimiza a recuperação de informações por meio de uma estrutura hierárquica de *tags* médicas. Diferente da recuperação linear convencional, o *U-Retrieval* opera em um fluxo que descende do global ao local: o sistema primeiro identifica entidades em níveis macro (camadas superiores do grafo) para filtrar o domínio relevante e, em seguida, acessa o nível granular das triplas para

extrair evidências precisas. Esse processo permite capturar tanto a visão geral da patologia quanto os detalhes clínicos específicos, equilibrando eficiência com um contexto abrangente. Avaliações utilizando *benchmarks* de perguntas médicas e avaliações clínicas demonstraram a superioridade do *MedGraphRAG* em relação a métodos tradicionais e *LLMs* específicos do domínio, estabelecendo um novo estado da arte em respostas baseadas em evidências na área da saúde.

Na mesma linha de *GraphRAG*, Schiffman (2024) apresentou uma prova de conceito para o emprego da *GraphRAG* com dados *FHIR* como alternativa à *RAG* tradicional [10]. O autor destaca a adequação dos grafos para gerenciar a natureza interconectada dos dados *FHIR*, facilitando ligações entre recursos e oferecendo uma visão mais holística e detalhada do contexto clínico. Essa abordagem também considera a dimensão temporal dos eventos médicos, como a análise da eficácia do tratamento ao longo do tempo, permitindo respostas mais detalhadas e contextuais em cenários de saúde complexos.

Para além da saúde, a *GraphRAG* tem sido aplicado em domínios mais amplos, demonstrando sua versatilidade e eficácia. Edge et al. (2024) foram pioneiros na *GraphRAG* para a sumarização global de grandes conjuntos de dados utilizando *KGs* derivados de *LLMs* [18]. Distanciando-se das estruturas tradicionais de recuperação de índices de grafos, seu estudo enfatiza a modularidade inerente dos grafos e a capacidade dos algoritmos de detecção de comunidades para particioná-los em módulos de comunidades relacionadas. Os resumos gerados por *LLMs* para essas comunidades cobrem de forma abrangente o índice do grafo e os documentos de entrada. Utilizando uma abordagem *map-reduce*, cada resumo da comunidade responde independentemente às consultas, que são então consolidadas em uma resposta global. Avaliações utilizando questões de conjuntos de dados reais revelaram que os métodos globais superam o *RAG* tradicional em compreensividade e diversidade, conforme avaliado por técnicas de avaliação de *LLMs* [35, 36]. A *GraphRAG*, particularmente com resumos de comunidades em níveis intermediários e baixos, demonstra desempenho superior na compreensão de questões amplas e temas diversos. No entanto, isso acarreta custos mais altos de *tokens*.

Diversos estudos exploraram o uso de grafos para aprimorar a qualidade dos dados em contextos de *LLMs*, empregando várias estratégias para enriquecer o contexto e melhorar a recuperação. Com base nisso, a estratégia *k-hops* tem sido amplamente utilizada para adicionar informações contextuais relevantes ao nó alvo, como evidenciado por vários estudos [4, 37, 38, 39, 40, 41, 42].

Por outro lado, Guo et al. (2023) empregam a busca em profundidade (*Depth-First Search*) em etapas específicas do pipeline [4], enquanto Dehghan et al. (2024) adotam o *Beam Search* como técnica de busca [43]. Essas abordagens destacam a flexibilidade e adaptabilidade dos métodos baseados em grafos para aprimorar o desempenho de *LLMs*

em diversas aplicações, solidificando ainda mais seu papel como uma ferramenta crítica na evolução dos sistemas *RAG*.

Os *LLMs* demonstraram potencial no processamento de dados de saúde estruturados, especialmente quando integrados ao padrão *FHIR*. Schmiedmayer et al. (2024) exploraram o uso de *LLMs* com dados *FHIR* estruturados para desenvolver um *chatbot* integrado a um aplicativo móvel [3]. Este *chatbot* emprega uma abordagem *RAG* utilizando busca lexical para aprimorar o acesso e a compreensão de informações de saúde pelos pacientes, com foco na democratização do acesso a dados de saúde e na melhoria da literacia em saúde. Seu estudo serve como referência fundamental para este trabalho, que utiliza o mesmo conjunto de dados para avaliar a qualidade da solução, enfatizando a democratização do acesso à informação em saúde.

De forma semelhante, *KMS Healthcare Innovation Labs* (2023) desenvolveu um produto utilizando *RAG* com busca por similaridade para extrair informações relevantes de documentos médicos eletrônicos em formato *FHIR* [11]. Seu trabalho aproveitou os *LLMs* para avaliar a qualidade e a precisão das respostas em aproximadamente 200 perguntas geradas pelo modelo, alcançando uma taxa de acerto de 86,3%. Isso ressalta a credibilidade e a viabilidade de sua abordagem para aplicações práticas na saúde.

A pesquisa e o desenvolvimento de sistemas de IA na saúde dependem fundamentalmente de conjuntos de dados robustos e representativos. Entre os mais proeminentes está o ecossistema MIMIC (*Medical Information Mart for Intensive Care*), incluindo suas versões MIMIC-III[44]. e MIMIC-IV [45]. Essas bases de dados, publicamente disponíveis, contêm uma vasta quantidade de dados clínicos anonimizados de pacientes de unidades de terapia intensiva, abrangendo notas clínicas, resultados de exames, diagnósticos e procedimentos. A riqueza e a complexidade do MIMIC o tornam um recurso padrão para treinar e avaliar modelos em tarefas como sumarização de prontuários eletrônicos, resposta a perguntas clínicas e extração de informações, sendo um pilar em múltiplos estudos citados neste trabalho.

Para além dos dados brutos de prontuários, a comunidade científica desenvolveu uma série de *benchmarks* e corpora especializados para avaliar rigorosamente o desempenho dos modelos. O conjunto MultiMedQA<sup>1</sup>, por exemplo, agrega diversos *datasets* de múltipla escolha como MedQA e PubMedQA, testando o conhecimento e o raciocínio médico dos *LLMs*. Para tarefas de verificação de fatos, são utilizados conjuntos de dados como FakeHealth[46] e PubHealth[47], que ajudam a medir a capacidade dos modelos de distinguir informações verdadeiras de desinformação. Adicionalmente, grandes repositórios de conhecimento, como o MedC-K [48], que contém milhões de artigos biomédicos e livros didáticos, e grafos de conhecimento como o UMLS [49], são cruciais para fundamentar as

---

<sup>1</sup><https://huggingface.co/collections/openlifescienceai/multimedqa-66098a5b280539974cefe485>

respostas dos modelos em fontes autorizadas, especialmente em arquiteturas avançadas como o *MedGraphRAG* [16].

Devido aos desafios de privacidade e acesso a dados de saúde reais, o uso de dados sintéticos e institucionais também se tornou uma prática comum. Ferramentas como o gerador Synthea™, que produz o conjunto de dados SyntheticMass, permitem criar registros de pacientes realistas, mas artificiais, viabilizando o desenvolvimento e a avaliação de soluções em um ambiente controlado, como demonstrado por Schmiedmayer et al. (2024)[3].

Em contraste, outros estudos, como o de Zaretsky et al. (2024), optam por utilizar dados proprietários de suas próprias instituições de saúde, como um conjunto específico de resumos de alta hospitalar, para garantir que a solução seja perfeitamente adaptada ao seu contexto clínico específico [14]. Essa diversidade de fontes de dados reflete as diferentes necessidades de validação, desde a avaliação generalizada em *benchmarks* públicos até a aplicação prática em cenários clínicos do mundo real.

A avaliação de sistemas de IA generativa na área da saúde emprega um espectro diversificado de métricas, que vão desde *benchmarks* quantitativos estabelecidos até avaliações qualitativas por especialistas. Métricas tradicionais de PLN formam a base para muitas tarefas. O ROUGE (Recall-Oriented Understudy for Gisting Evaluation) e o BLEU (Bilingual Evaluation Understudy) são frequentemente utilizados para avaliar a qualidade de resumos, medindo a sobreposição de n-gramas entre o texto gerado e um texto de referência [13, 15]. Para uma análise mais semântica, métricas como o *BERTScore*, que compara a similaridade contextual dos *embeddings* das palavras, oferecem uma avaliação mais robusta da fidelidade do conteúdo. Em tarefas de classificação ou resposta a perguntas com respostas diretas, métricas como Acurácia, Precisão, Revocação e F1-Score são padrão para medir o desempenho do modelo em *benchmarks* estabelecidos, como os do conjunto de MultiMedQA<sup>2</sup> [16].

Em arquiteturas de *RAG* e *GraphRAG*, a avaliação pode ser especializada em duas frentes: a qualidade do componente de recuperação e a qualidade da geração final. Para o módulo de recuperação, o objetivo é medir a eficácia com que os documentos ou nós de grafo relevantes são trazidos para o contexto. Métricas padrão de Recuperação de Informação (RI) são aplicadas, como *Recall@K*, que mede a proporção de vezes que um documento relevante é encontrado entre os K melhores resultados, e o *Mean Reciprocal Rank* (MRR), que avalia a posição do primeiro resultado relevante, penalizando rankings onde a resposta correta aparece mais abaixo na lista [40]. A métrica *Hits@K* é também comumente utilizada, indicando se a resposta correta está presente no top-K dos itens recuperados [37].

---

<sup>2</sup><https://huggingface.co/datasets/openlifescienceai/multimedqa>

Devido à natureza crítica e de alto risco do domínio da saúde, a avaliação puramente quantitativa é muitas vezes insuficiente. Por isso, a avaliação por especialistas humanos (*Subject-Matter Experts*, SMEs) é considerada o padrão-ouro. Estudos frequentemente recorrem a equipes de médicos ou clínicos para julgar as saídas dos modelos com base em critérios que são difíceis de automatizar, como precisão clínica, completude da informação, relevância para o caso e ausência de alucinações [3]. Além disso, para aplicações voltadas ao paciente, métricas como legibilidade e compreensibilidade são cruciais para garantir que a informação gerada seja útil e acessível para um público não especializado, equilibrando simplicidade com a manutenção da acurácia médica [14].

Para superar o alto custo e a baixa escalabilidade da avaliação humana, uma abordagem emergente e poderosa é o uso de *LLMs* como juízes automatizados. Nessa metodologia, um modelo de ponta, recebe a pergunta, a resposta gerada pelo sistema em teste e, opcionalmente, o contexto de referência, sendo instruído a avaliar a resposta com base em um conjunto de critérios qualitativos. Métricas como Compreensividade, Diversidade, Empoderamento e Direcionamento podem ser avaliadas dessa forma [18]. Outras implementações criam sistemas de teste automatizados que geram perguntas e validam as respostas, retornando um veredito binário para extrair acurácia das repostas, agilizando drasticamente o ciclo de avaliação [11].

# Capítulo 4

## Proposta

Este capítulo apresenta a metodologia proposta para o desenvolvimento deste projeto, que visa construir uma *GraphRAG* com o objetivo de superar as limitações em dados médicos interconectados comparado aos métodos de *RAG* tradicionais. O objetivo principal é desenvolver um sistema de diálogo, fundamentado nos trabalhos de Schriedmayer et al. (2024), *KMS Healthcare Innovation Labs* (2023) e Schifman (2024), que facilite a compreensão e o acesso dos pacientes aos seus dados médicos [3, 11, 10]. Este sistema é projetado para sumarizar informações de registros médicos eletrônicos no padrão *FHIR*, utilizando uma base de conhecimento em grafos como principal suporte para a recuperação e contextualização da informação.

A avaliação do sistema é realizada por meio da comparação entre métodos bases de busca lexical e por similaridade, incluindo suas variantes com expansão via grafos. São empregados métricas de avaliação de qualitativas como *Answer Relevancy*, *Contextual Relevancy* e *Hallucination* [35, 36]. Adicionalmente, métricas subjetivas como *Comprehensiveness*, *Diversity*, *Empowerment* e *Directness* [18] são consideradas. O desempenho é analisado em cinco diferentes modelos de *LLMs*, com medições de quantidade de *tokens* e tempo de resposta, visando aferir a viabilidade do sistema.

O restante deste capítulo é dividido em quatro seções: na Seção 4.1 apresenta-se o conjunto de dados utilizados no projeto; na Seção 4.2 detalha-se o processo para transformação dos dados em uma base de grafos e as principais etapas de pré-processamento; a Seção 4.3 descreve a estruturação do sistema *GraphRAG*, as técnicas de recuperação de informação e os experimentos e na Seção 4.4 abordam-se em detalhes as métricas e o processo de avaliação utilizados no projeto

## 4.1 Base de Dados

Nesta seção, apresentamos a base de dados empregada neste estudo, detalhando sua origem, as características dos pacientes sintéticos e a justificativa para as escolhas metodológicas relacionadas à sua seleção e ao seu processamento.

Este estudo utiliza um conjunto de dados médicos sintéticos de cinco pacientes, extraídos do trabalho de Schmiedmayer et al. (2024) [3]. Os dados foram originalmente gerados por meio do *Synthea*<sup>1</sup>, uma ferramenta de código aberto amplamente reconhecida por sua capacidade de criar prontuários de saúde sintéticos e realistas. Esses dados são provenientes do conjunto de dados público *SyntheticMass (Versão 2)*<sup>2</sup>.

A seleção dos cinco pacientes do conjunto de dados *SyntheticMass* foi criteriosa e feita pelo trabalho Schmiedmayer et al. (2024) [3], visando a uma avaliação reprodutível e imparcial das respostas do modelo de linguagem. Para garantir a representatividade, o conjunto de dados completo foi particionado em dez categorias distintas, cada uma correspondendo a uma condição ou procedimento cardiovascular específico, alinhando-se à especialidade da maioria dos médicos especialistas que avaliaram o estudo original. Após essa partição e a exclusão de pacientes que não se enquadravam nos critérios, foi selecionado um grupo balanceado, assegurando que todos os pacientes estivessem vivos, que pelo menos dois deles tivessem alergias registradas e que houvesse um equilíbrio representativo em termos de gênero autorrelatado, origem étnica e faixas etárias, que variavam de 8 a 82 anos.

A Tabela 4.1 resume as principais características demográficas e médicas desses cinco pacientes, incluindo idade, gênero, condições primárias, alergias documentadas e medicações em uso. Esta visão geral demonstra a diversidade dos perfis dos pacientes no conjunto de dados, abrangendo uma ampla gama de idades e históricos médicos complexos.

É importante ressaltar que não há certeza se os LLMs utilizados no presente trabalho tiveram contato prévio com esses dados médicos durante seu treinamento. O estudo utiliza os modelos descritos na Subseção 4.3.4. As datas de corte de conhecimento para os diferentes modelos precederam a publicação do estudo de referencia. Assim sendo, como o conjunto de dados *SyntheticMass* é público e foi gerado antes dessas datas, é plausível que exista uma possibilidade que os dados médicos tenham sido incluído nos dados de treinamento de algumas versões do modelo, o que poderia influenciar a performance do LLM na tarefa de interpretação dos prontuários.

---

<sup>1</sup><https://synthetichealth.github.io/synthea/>

<sup>2</sup><https://github.com/StanfordBDHG/LLMonFHIR/issues/50>

Tabela 4.1: Resumo dos dados dos cinco pacientes selecionados do trabalho de Schmiedmayer et al. (2024) [3]

Nome	Idade	Gênero	Condições	Alergias	Medicações
Beatris270 Bogan287	8	F	Aortic valve stenosis (disorder) Perennial allergic rhinitis Atopic dermatitis	Latex (substance) Bee venom (substance) Mold (organism) House dust mite (organism) Animal dander (substance) Grass pollen (substance) Tree pollen (substance) Aspirin	Fexofenadine hydrochloride 30 MG Oral Tablet Epinephrine 1 MG/ML Auto-Injector 0.3 ML
Milton509 Ortiz186	26	M	Hypertension Hypoxemia (disorder) Stress (finding)	-	MedicationRequest amlODIPine 2.5 MG Oral Tablet
Edythe31 McDermott739	49	F	Body mass index 30+- obesity (finding) Received higher education (finding) Prediabetes Anemia (disorder) Victim of intimate partner abuse (finding) Cardiac Arrest History of cardiac arrest (situation)	-	Jolivetite 28 Day Pack
Jacklyn830 Veum823	72	F	Essential hypertension (disorder) Miscarriage in first trimester Ischemic heart disease (disorder) Chronic kidney disease stage 3 (disorder) Proteinuria due to type 2 diabetes mellitus (disorder) Social isolation (finding) Sprain (morphologic abnormality)	-	Nitroglycerin 0.4 MG/ACTUAT Mucosal Spray Simvastatin 20MG Oral Tablet Clopidogrel 75 MG Oral Tablet 24 HR metoprolol succinate 100 MG Extended Release Oral Tablet Acetaminophen 325 MG Oral Tablet Hydrochlorothiazide 25 MG Oral Tablet
Allen332 Ferry570	82	M	Chronic sinusitis (disorder) Hypertension Served in armed forces (finding) Received higher education (finding) Body mass index 30+- obesity (finding) Prediabetes Anemia (disorder) Opioid abuse (disorder) Atrial Fibrillation Neoplasia of prostate Carcinoma in situ of prostate (disorder) Chronic intractable migraine without aura Victim of intimate partner abuse (finding) Stress (finding) Alzheimer's disease (disorder)	Animal dander (substance) Penicillin V Peanut (substance)	Galantamine 4 MG Oral Tablet Warfarin Sodium 5 MG Oral Tablet doxycycline hyclate 100 MG 1 ML DOCEtaxel 20 MG/ML Injection 0.25 ML Leuprolide Acetate 30 MG/ML Prefilled Syringe lisinopril 10 MG Oral Tablet Verapamil Hydrochloride 40 MG Digoxin 0.125 MG Oral Tablet

A Tabela 4.2 detalha adicionalmente a distribuição dos diferentes tipos de recursos *FHIR* para esses cinco pacientes. apresentando as contagens para as principais categorias de recursos, como Condições, Observações, Encontros, Requisições de Medicação, Procedimentos, Relatórios de Diagnóstico, Reivindicações e Explicações de Faturas, juntamente com o número total de recursos *FHIR* para cada paciente. Esta tabela evidencia a riqueza e a interconexão dos dados médicos, destacando a necessidade de mecanismos de recuperação avançados para navegar e extrair informações relevantes de forma eficaz [1].

Tabela 4.2: Contagem dos Principais Tipos de Recursos e Total de Recursos *FHIR* por Paciente

Paciente	Condição	Observação	Encontro	RequisiçãoMed.	Procedimento	RelatórioDiag.	Reivindicação	ExplicaçãoFatura	Total
Allen332	63	252	83	191	63	124	273	82	1235
Beatris270	15	233	31	6	18	33	37	37	500
Edythe31	28	132	35	16	34	55	51	35	442
Jacklyn830	69	1488	56	47	217	243	103	103	2405
Milton509	28	705	19	11	92	104	30	19	1129

Tabela 4.3: Relação entre a quantidade de *tokens* e caracteres no arquivo de cada paciente.

Paciente	Qtde de Recursos <i>FHIR</i>	Total <i>Tokens</i>	Total Caracteres
Beatris270 Bogan287	500	341.053	913.142
Milton509 Ortiz186	1.129	662.747	1.805.901
Edythe31 Morar593	442	359.113	930.922
Jacklyn830 Veum823	2.405	1.738.007	4.809.609
Allen322 Ferry570	1.235	908.412	2.358.050

Ademais, a escolha pelo padrão *FHIR* justifica-se por sua crescente prevalência e padronização na indústria da saúde, promovendo a interoperabilidade e a troca de dados [6]. A estrutura e as restrições bem definidas do *FHIR* fornecem uma base robusta para a construção de sistemas avançados de recuperação de informação.

Os dados possuem um volume de dados que excede significativamente a janela de contexto dos modelos *LLMs* mais avançados, como detalhado na Tabela 4.3. Essa característica reforça a necessidade de adotar uma abordagem de *RAG*, permitindo gerenciar o contexto de forma mais eficiente e garantindo uma interação adequada com os modelos. A utilização da *RAG* é crucial para lidar com a complexidade e volume dos dados *FHIR*, possibilitando a recuperação precisa de informações relevantes para alimentar os *LLMs*.

## 4.2 Construção da base de conhecimentos em grafos

Como abordado na revisão teórica do capítulo *GraphRAG* e destacado pela literatura [2], uma das etapas fundamentais em uma *GraphRAG* é a construção da base de conhecimentos em grafos. Essa etapa é crucial para estabelecer os relacionamentos entre as informações, que são utilizados para fornecer contexto aos modelos de linguagem. Os relacionamentos criados no grafo permitem expandir a busca por informações relevantes.

Nesta etapa, foi desenvolvida uma base de conhecimento em grafos com base na abordagem proposta no trabalho de Schiffman (2024) [10]. O objetivo principal é converter os dados em formato *JSON* de arquivos *FHIR* para uma estrutura em grafos, estabelecendo relacionamentos entre os recursos associados. O fluxo do processo está ilustrado na Figura 4.1, as etapas chave incluem: primeiro, a adição de metadados para distinguir e rastrear pacientes individualmente na KG; segundo, é feito um processo para extrair recursos *FHIR* e suas referências, formando os nós e arestas do grafo.

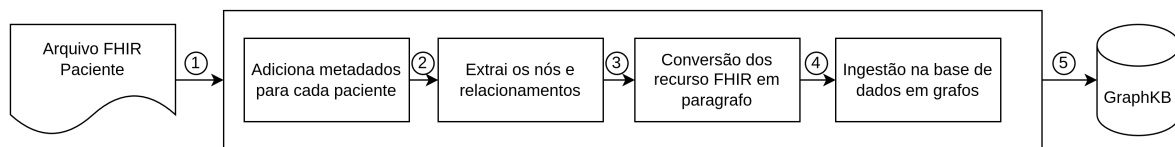


Figura 4.1: Processo de geração da base de conhecimentos em grafos.

Para otimizar o uso de *tokens* e melhorar a qualidade das respostas dos modelos de linguagem, os recursos *FHIR* passam por um pré-processamento inspirado no pré-processamento feito no trabalho da *KMS Healthcare Innovation Labs* (2023) [11]. Essa conversão facilita a utilização dos dados para os modelos de LLMs, conforme demonstrado em trabalhos de referência [50], que evidenciou uma melhoria na acurácia dos resultados quando os dados são representados de forma textual estruturada.

A Tabela 4.4 apresenta um exemplo comparativo de um recurso *FHIR* em sua forma bruta e após o pré-processamento.

### 4.2.1 Principais Etapas do Pré-processamento

O processo de pré-processamento é executado de forma modular e paralela, garantindo eficiência na manipulação de grandes volumes de dados. As etapas principais são:

1. **Ingestão e Desagregação dos Recursos FHIR:** O fluxo de processamento inicia-se com a leitura de um arquivo *JSON* que encapsula um pacote de recursos *FHIR* (*Bundle*). Este grupo é uma coleção de recursos individuais, como *Patient*,

Tabela 4.4: Comparação entre recurso *FHIR* bruto e pré-processado. Na esquerda o recurso *FHIR* bruto, na direita o recurso *FHIR* pré processado. A instrução utilizada nessa etapa pode ser encontrada no Apêndice B

Exemplo de recurso <i>FHIR</i> Bruto	Respectivo recurso <i>FHIR</i> convertido
<pre>{ cf18362b-5057-8a39-039b-dbc9e13ed518, resource: {"resourceType": "MedicationRequest", meta: {"profile": ["http://hl7.org/fhir/us/core /StructureDefinition/us-core-medicationrequest"]} }, status: "active", intent: "order", category: [{"coding": [ {"system": "http://terminology.hl7.org/CodeSystem /medicationrequest-category", "code": "community", "display": "Community"}], "text": "Community"}], medicationCodeableConcept: {"coding": [{"system": "http://www.nlm.nih.gov/research/umls/rxnorm", "code": "1870230", "display": "NDA020800 0.3 ML Epinephrine 1 MG/ML Auto-Injector"}], "text": "NDA020800 0.3 ML Epinephrine 1 MG/ML Auto-Injector" }}, subject: {"reference": "urn:uuid:5b3645de-a2d0-d016-0839-bab3757c4c58"}}, encounter: {"reference": "urn:uuid:0bc7f36f-9d73-bdef-bfaa-b097c4b99dbc"}}, authoredOn: "2017-08-30T12:14:48+00:00", requester: {"reference": "Practitioner?identifier= http://hl7.org/fhir/sid/us-npi 9999990697", "display": "Dr. Alvin56 Crona259"}}, dosageInstruction: [{"sequence": 1, "text": "Take as needed.", "asNeededBoolean": true}] }</pre>	<pre>Resource TYpe: Medication Request Status: Active Intent: Order Category: Community (Medical Coding Classification System: http://terminology.hl7.org /CodeSystem/medicationrequest-category) Medication: NDA020800 0.3 ML Epinephrine 1 MG/ML Auto-Injector (Medical Coding Classification System: http://www.nlm.nih.gov/research/umls/rxnorm) Authored On: August 30, 2017, 12:14:48 PM +00:00 Requester: Dr. Alvin56 Crona259 Dosage Instruction: Take as needed. Dosage Instruction Sequence: 1 Dosage Instruction As Needed: True</pre>

**Encounter, Condition e Observation.** O *script* de ingestão desagrega o arquivo, tratando cada recurso como uma unidade independente de processamento. Para otimizar o desempenho, essa desagregação é executada de forma paralela, onde múltiplos recursos são processados simultaneamente em diferentes *threads*.

**2. Conversão de Recursos para Nós e Arestas de Grafo:** Para cada recurso *FHIR*, a estrutura hierárquica do JSON é decomposta que compõem um grafo: nós e arestas. Essa etapa utiliza *scripts* de referencia do trabalho Schifman (2024), com melhorias em processamento paralelo [10].

- **Nós:** O recurso *FHIR* é convertido em um nó. Para facilitar a indexação e a busca, a estrutura aninhada do JSON é normalizada em um conjunto de pares chave-valor, que se tornam as propriedades do nó. O tipo de recurso *FHIR* (*resourceType*) é utilizado como o rótulo do nó no grafo (por exemplo, *MedicationRequest*, *Patient*). No exemplo da Tabela 4.4 esse valor é *MedicationRequest*.
- **Arestas (Edges):** As relações entre os recursos são identificadas e convertidas em arestas. O padrão *FHIR* utiliza campos como *reference* para apontar para outros recursos, estabelecendo conexões lógicas. Por exemplo, um recurso *MedicationRequest* pode ter uma referência para um *Encounter*, *Practitioner* e um *Patient*, resultando em arestas que conectam os respec-

tivos nós no grafo, representando as relações “ocorrido durante”, “requisitado pelo” e “afeta o paciente”. No exemplo da Tabela 4.4 esse recurso possui algumas referencias como: `subject`, `encounter` e `requester` essas referencias referenciam IDs de outros recursos, esses relacionamos são exemplificados pela Figura 4.2.

3. **Enriquecimento Semântico dos Dados:** Para além da estruturação, o processo enriquece os dados com informações que facilitam buscas e análises semânticas. Esta é uma das etapas mais inovadoras da metodologia proposta:

- **Conversão para texto estruturado:** Utilizando um modelo de LLM, o JSON de cada recurso *FHIR* é traduzido para uma representação textual simples e legível por humanos. Essa técnica transforma dados técnicos e codificados em uma descrição clara e concisa, preservando as informações clínicas essenciais. Por exemplo, um JSON de uma `MedicationRequest` se torna uma frase como exemplificado na Tabela 4.4, nessa etapa é utilizado a instrução apresentada no Apêndice B.
- **Geração de vetores:** A representação textual gerada é então convertida em um vetor por meio de um modelo de linguagem de vetores. Este vetor numérico captura o significado semântico do texto, permitindo que o banco de dados do grafo realize buscas de similaridade. Com isso, é possível encontrar recursos clinicamente semelhantes, mesmo que não compartilhem os mesmos códigos de diagnóstico ou termos. Nessa etapa foi utilizado o modelo de vetorização BAAI/bge-small-en-v1.5<sup>3</sup> que possui 33.4 Milhões de parâmetros, nas configurações padrões, que por sua simplicidade e performance competitiva no idioma em inglês pode ser utilizado localmente com processamento paralelo resultando em um vetor de 384 dimensões.

4. **Persistência de Dados no Banco de Grafo:** Após a transformação e o enriquecimento, os nós e arestas gerados são persistidos no banco de dados. Este processo envolve a formulação de consultas *Cypher* que criam os nós com suas respectivas propriedades (incluindo o texto e o vetor) e as arestas que definem as relações entre eles. Adicionalmente, são criados índices vetoriais no banco de dados para otimizar as buscas por similaridade semântica, utilizando os vetores gerados na etapa anterior. A Figura 4.3 exemplifica como esse nó é representando com as propriedades.

Essa metodologia de pré-processamento garante que os dados clínicos, que são inerentemente complexos, sejam transformados em uma estrutura de grafo rica e interconectada,

---

<sup>3</sup><https://huggingface.co/BAAI/bge-small-en>

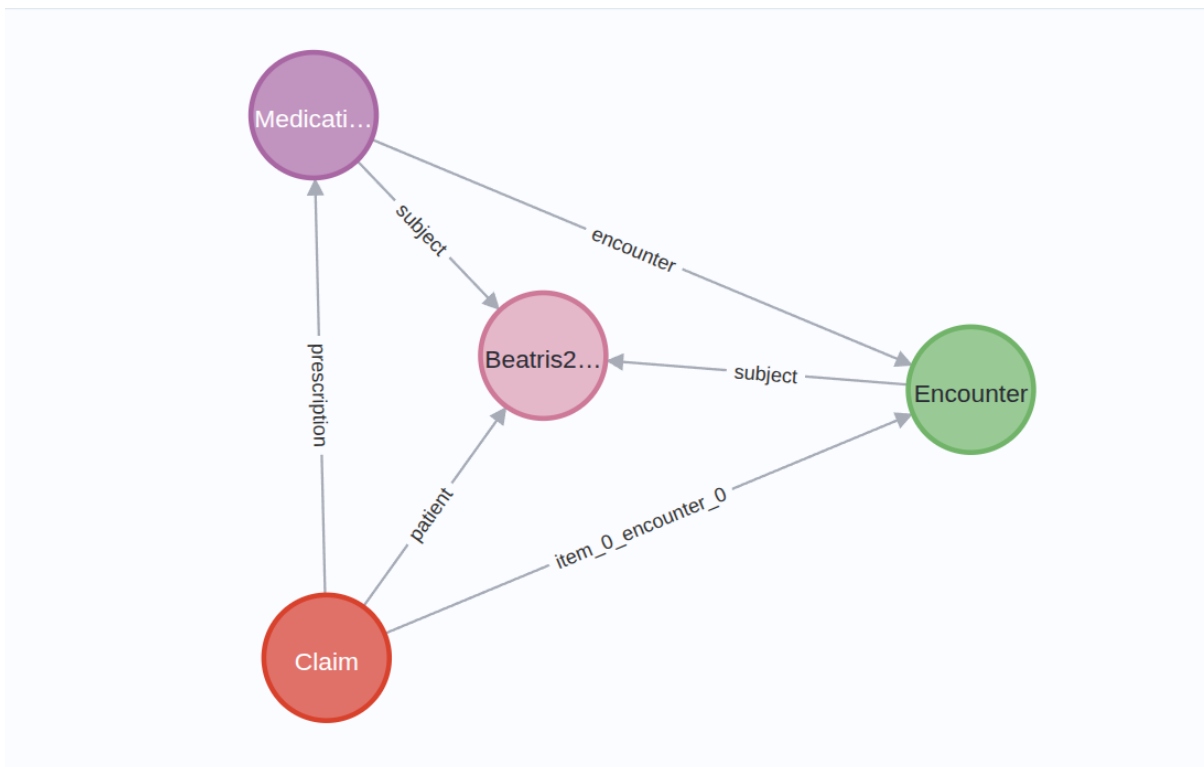


Figura 4.2: Exemplo dos relacionamos de um nó em específico.

Propriedade	Valor
<b>raw_fhir</b>	{ "resourceType": "MedicationRequest", "id": "cf18362b-5057-8a39-039b-dbc9e13ed518", "meta": { "profile": [ "http://hl7.org/fhir/us/core/StructureDefinition/us-core-medicationrequest" ] }, "status": "active", "intent": "order", "category": [ { "coding": [ { "system": "http://terminology.hl7.org/CodeSystem/me..." } ] } ] } <a href="#">Show all</a>
<b>requester_display</b>	Dr. Alvin56 Crona259
<b>requester_reference</b>	Practitioner?identifier=http://hl7.org/fhir/us-npi 9999990697
<b>resource_type</b>	MedicationRequest
<b>status</b>	active
<b>subject_reference</b>	urn:uuid:5b3645de-a2d0-d016-0839-bab3757c4c58
<b>text</b>	Resource Type: Medication Request Status: Active Intent: Order Category: Community (Medical number: http://terminology.hl7.org/CodeSystem/medicationrequest-category code: community) Medication: NDA020800 0.3 ML Epinephrine 1 MG/ML Auto-Injector (Medical number: http://www.nlm.nih.gov/research/umls/r...

Figura 4.3: Exemplo do nó persistido no banco de dados, onde na esquerda é possível ver as propriedades.

pronta para ser explorada para fins de pesquisa, análise clínica e suporte à decisão. A combinação de dados estruturados com representações semânticas abre novas possibilidades para a descoberta de padrões e a compreensão holística de prontuários médicos.

### 4.3 Sistema de diálogo

Este projeto propõe a construção de um sistema de diálogo baseado em *GraphRAG*, que utiliza um grafo como base de conhecimento para responder às perguntas dos usuários. A arquitetura geral do sistema está representada na Figura 4.4. Nas subseções a seguir serão apresentadas cada etapa dessa arquitetura proposta.

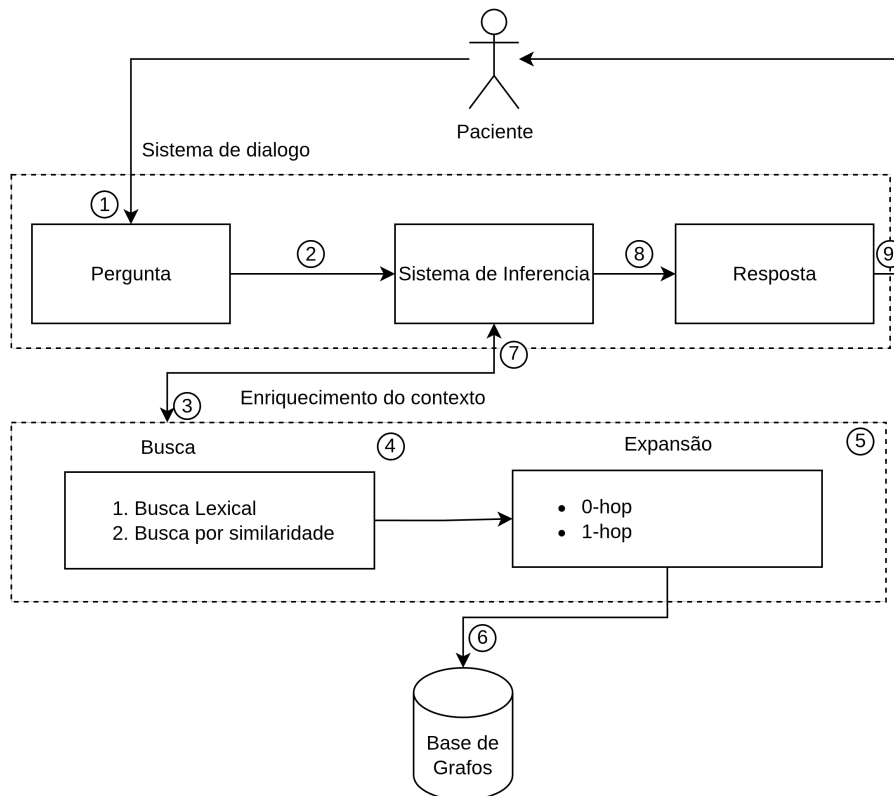


Figura 4.4: Arquitetura geral do sistema de diálogo proposto. À esquerda, estão os componentes de alto nível da arquitetura.

#### 4.3.1 Fluxo da arquitetura

Conforme ilustrado na Figura 4.4, o fluxo de funcionamento da arquitetura pode ser explicado passo a passo:

1. **Estímulo de Entrada:** O processo inicia com um estímulo de entrada na forma de uma pergunta (*prompt*).
2. **Análise pela LLM:** A pergunta é então analisada por um *LLM*.
3. **Busca de Conhecimento Adicional:** Identifica-se a necessidade de buscar conhecimentos adicionais para responder à pergunta inicial.

4. **Seleção da Estratégia de Busca:** Seleciona-se o tipo de busca previamente configurado:

- **Busca por Similaridade:** Se esta for a opção, a pergunta inicial é vetorizada.
- **Busca Lexical:** Realiza-se a extração de possíveis tipos de recursos *FHIRs* que podem responder à pergunta inicial.

Em seguida, a busca é realizada na base de conhecimento de grafos.

5. **Expansão de Nós (Opcional):** Caso configurada, ocorre a expansão dos nós adjacentes aos nós inicialmente retornados na etapa anterior (expansão *k-hops*).

6. **Consulta ao Banco de Dados em Grafos:** Os parâmetros são enviados ao banco de dados em grafos, e os nós correspondentes são retornados em formato de texto.

7. **Formatação e Inserção no Modelo:** Os nós retornados são formatados em uma *adjacency list* e passados para o modelo, juntamente com a instrução de sistema e a pergunta inicial.

8. **Inferência da Resposta:** O modelo realiza a inferência da resposta, utilizando as instruções, a pergunta e o contexto fornecido.

9. **Apresentação da Resposta:** A resposta final é então apresentada ao usuário.

### 4.3.2 Recuperação de Informação

Na etapa de recuperação de informações, são empregadas duas estratégias utilizadas em sistemas de recuperação inspiradas nos trabalhos de Schmiedmayer et al. (2024), *KMS Healthcare Innovation Labs* (2023) e Schifman (2024) [3, 11, 10]:

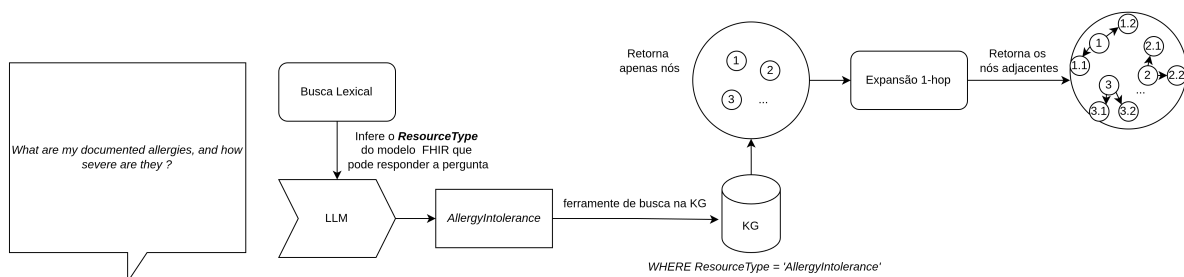


Figura 4.5: Representação gráfica das etapas da Busca Lexical.

- **Busca Lexical:** A Figura 4.5 ilustra essa etapa. Baseada nos conhecimentos internos das *LLMs* sobre o padrão *FHIR*, esta abordagem infere, a partir da pergunta de

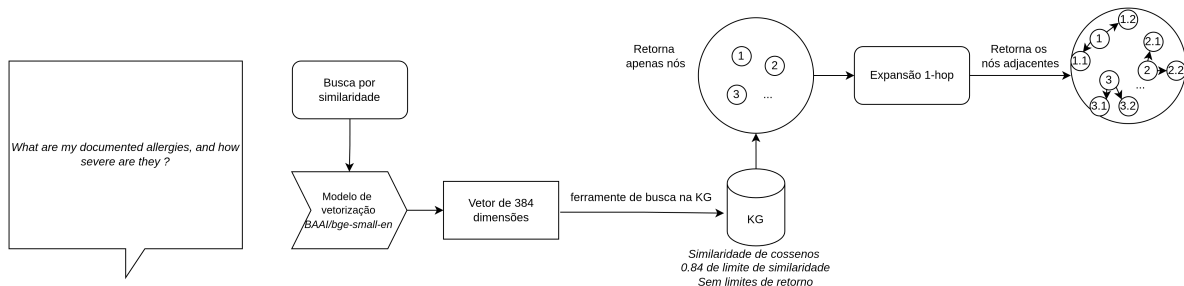


Figura 4.6: Representação gráfica das etapas da Busca por Similaridade.

entrada, a chave-valor do tipo de recurso *FHIR* correspondente para a busca no banco de dados. Essa técnica foi detalhada no trabalho Schmiedmayer et al. (2024) [3]. Importante notar que não existe limite de nós de retorno.

- **Busca por Similaridade:** A Figura 4.6 mostra essas etapas. Conforme descrito nos trabalhos [11, 15, 17], essa busca realiza a representação da pergunta de entrada e dos textos no banco de dados em vetores, comparando-os para identificar correspondências relevantes. Nessa etapa são utilizados os parâmetros mostrados na Apêndice D, com índice de similaridade de 84%, sem limite de nós de retorno e similaridade de cossenos para fazer a comparação entre os vetores.

De acordo com os estudos apresentados na revisão teórica da seção 2.3, *GraphRAG*, a etapa de recuperação de informações inclui uma expansão da busca para capturar relações que, em buscas tradicionais, seriam ignoradas. Neste trabalho, optou-se por explorar uma técnica paramétrica conhecida como expansão por nós adjacentes ou *k-hops expansion*, conforme abordado extensivamente na literatura [4, 37, 38, 39, 40, 41, 42]. Essa técnica amplia a busca para incluir os  $k$  nós adjacentes ao nó inicialmente recuperado.

### 4.3.3 Formatação do grafo recuperado

Como apresentado na Subseção 2.3, uma das etapas da *GraphRAG* é a formatação do sub-grafo retornado na etapa de expansão para um formato de texto que seja interpretado pelas LLMs. Conforme abordado na Seção 4.3, o grafo retornado nesta etapa representado no formato *Adjacency List*, considerando que o trabalho do Guo et al. (2023) aponta desafios no uso de grafos em modelos *LLM* [4], contudo destaca a eficácia do formato *Adjacency List*, *Edge List*, *GraphML* (*Graph Markup Language*) e *GML* (*Graph Modeling Language*), esses formatos de representação de grafos podem ser visualizados na Tabela 4.5.

Para o presente estudo, adotou-se o formato *Adjacency List* por ser menos verboso em comparação ao demais formatos e por não necessitar especificar direcionalidade dos relacionamentos. A Tabela 4.6 mostra a formatação escolhida.

Tabela 4.5: Comparações entre Formatos de Representação de Grafo no trabalho Guo et al. (2023) [4]. Na esquerda o formato de representação, na direita o exemplo correspondente.

Formato de Representação	Exemplo para o Grafo
<i>Adjacency List</i>	1: 2, 3 2: 1, 3 3: 1, 2, 4 4: 3
<i>Edge List</i>	(1, 2) (1, 3) (2, 3) (3, 4)
<i>GraphML (Graph Markup Language)</i>	<pre>&lt;?xml version="1.0" encoding="UTF-8"?&gt; &lt;graphml xmlns="http://...graphdrawing.org/xmlns" &lt;graph id="G" edgedefault="undirected" &lt;node id="1"/&gt; &lt;node id="2"/&gt; &lt;node id="3"/&gt; &lt;node id="4"/&gt; &lt;edge source="1" target="2"/&gt; &lt;edge source="1" target="3"/&gt; &lt;edge source="2" target="3"/&gt; &lt;edge source="3" target="4"/&gt; &lt;/graph&gt; &lt;/graphml&gt;</pre>
<i>GML (Graph Modeling Language)</i>	<pre>graph [ comment "Um grafo de exemplo simples" directed 0 node [ id 1 ] node [ id 2 ] node [ id 3 ] node [ id 4 ] edge [ source 1 target 2 ] edge [ source 1 target 3 ] edge [ source 2 target 3 ] edge [ source 3 target 4 ] ]</pre>

Tabela 4.6: Formato de *Adjacency List* utilizado no presente trabalho.

Formato <i>Adjacency List</i>	Exemplo
{'Main health record': '<Dados do Nó>', 'Related health records': [ ( 'Relationship': '<Tipo de relacionamento>', 'Health Record' : '<Dados do Nó relacionado>'), ( 'Relationship': '<Tipo de relacionamento N>', 'Health Record' : '<Dados do Nó relacionado N>'), ]}	{'Main health record': 'Resource Type: Medication Request Status: Stopped Intent: Order ... Dosage Instruction As Needed: True , 'Related health records': [ ( 'Relationship': 'encounter', 'Health Record': 'Resource Type: Encounter Status: Finished Class: ambulatory Type: Encounter for symptoms ... Reason: Otitis media Location: HOLY FAMILY HOSPITAL Service Provider: HOLY FAMILY HOSPITAL ) ... ]}

Caso o grafo resultante exceda o limite da janela de contexto da *LLM*, é aplicada uma etapa de *truncamento*, na qual os textos que ultrapassarem a capacidade de contexto do modelo serão removidos. Dessa forma, mesmo com o contexto truncado, ele ainda poderá ser utilizado. Essa abordagem tem como objetivo garantir uma comparação justa entre os modelos, evitando o uso de técnicas de sumarização [18]. A aplicação de sumarização poderia comprometer a equidade da comparação entre diferentes modelos, uma vez que introduziria variações adicionais no processamento do conteúdo.

### 4.3.4 Experimentos

Para garantir consistência, os experimentos utilizaram as mesmas instruções, minimizando variações provenientes desses elementos. As seguintes metodologias de busca foram comparadas:

- Busca Lexical (L) e Busca por Similaridade (S);
- Variação na quantidade de *k-hops* (0 a 1) na expansão das buscas (L0, L1, S0 e S1);
- No conjunto de perguntas da Tabela 4.7.
- Para os cinco pacientes da Tabela 4.1

Os experimentos envolverão diversos modelos de *LLMs*, tanto proprietários quanto de código aberto, incluindo *GPT-4o-2024-08-06*<sup>4</sup>, *Gemini-1.5-pro*<sup>5</sup>, *Claude-v3-opus*<sup>6</sup>, *Llama-3-70b-instruct-awq*<sup>7</sup> e *DeepSeek-R1-Distill-Llama-70B*<sup>8</sup>.

Foram implementadas as estratégias de busca como busca lexical e busca por similaridade com e sem expansão de vizinhos adjacentes, ou seja, *0-hop* e *1-hop*. Para simplificar a nomenclatura, segue a conversão para as próximas seções e capítulos:

- Busca Lexical (L0): sem expansão de grafo - referencia do trabalho [3];
- Busca Lexical (L1): com expansão de *1-hop* no grafo;
- Busca por Similaridade (S0): sem expansão de grafo - referencia do trabalho [11];
- Busca por Similaridade (S1): com expansão de *1-hop* no grafo.

---

<sup>4</sup><https://platform.openai.com/docs/models/gpt-4o>

<sup>5</sup><https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-pro>

<sup>6</sup><https://www.anthropic.com/news/claude-3-family>

<sup>7</sup><https://huggingface.co/casperhansen/llama-3-70b-instruct-awq>

<sup>8</sup><https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B>

## 4.4 Avaliação

Para a avaliação do projeto, convencionou-se utilizar as perguntas Q1, Q3 e Q4 do trabalho de Schmiedmayer et al. (2024)[3], conforme apresentado na Tabela 4.7. Adicionando outras cinco perguntas com o objetivo de avaliar diferentes dimensões não exploradas no trabalho de referência como; explorar as alergias, equipes médicas, custos, procedimentos e imunizações dos pacientes.

Tabela 4.7: Perguntas avaliadas para cada paciente no presente trabalho.

ID	Pergunta
Q1 [3]	What are my current medications and how should I be taking them?
Q2	What are my documented allergies, and how severe are they ?
Q3 [3]	Can you summarize my current medical conditions ?
Q4 [3]	What are my recent laboratory values, what do they mean, and how can I improve them ?
Q5	Can you summarize my care plan history ?
Q6	Can you provide a breakdown of my medical bills ?
Q7	What procedures have I undergone recently, and what were the outcomes ?
Q8	Can you summarize my immunization history ?

As métricas de avaliação qualitativos adotados para a avaliação seguem as recomendações do *LLM-as-a-judge* [35, 36], conforme descrito na subseção 2.6.2 de Avaliação de RAG. As métricas incluem: *Answer Relevancy (ARS)*; *Contextual Relevancy (CRS)* e *Hallucination (HALS)*.

Além disso, serão analisados outros aspectos subjetivos mencionados nos trabalhos relacionados incluindo: *Comprehensiveness*; *Diversity*; *Empowerment*; *Directness* e *Overall winner* [13, 18, 32].

Essas métricas foram escolhidas devido a não necessidade de ter uma resposta esperada, como não existe uma base dados de respostas esperadas para cada pergunta. Onde nas métricas qualitativas são utilizadas as perguntas de entrada, o contexto fornecido e a resposta gerada pelo modelo para fazer o calculo de cada métrica. Por outro lado, nas métricas subjetivas é feita uma comparação direta entre a resposta gerada pelas técnicas, nessa etapa é feito comparação em pares de todas as estratégias utilizadas utilizando a instrução do Apêndice C.

Dada a inexistência de um respostas de referência para o conjunto de dados utilizado, a avaliação concentra-se primordialmente na etapa de geração de respostas. Essa decisão metodológica, alinhada a práticas observadas na literatura recente, justifica-se pela complexidade intrínseca à validação da etapa de recuperação no domínio clínico. Uma auditoria precisa da relevância do contexto recuperado exigiria a participação direta de especialistas em saúde para validar a acurácia médica das informações extraídas, o que foge ao escopo técnico deste trabalho. Assim, prioriza-se o uso de métricas baseadas em

*LLM-as-a-judge*, que permitem aferir a consistência e a utilidade da resposta final a partir do contexto fornecido.

A avaliação também inclui um estudo sobre a viabilidade técnica do sistema, Para isso, serão analisados o tempo de resposta e a quantidade total de *tokens* necessários para gerar as respostas na arquitetura *GraphRAG*. Esses dados permitirão avaliar o desempenho e a escalabilidade do sistema proposto, bem como sua viabilidade técnica em cenários práticos.

#### 4.4.1 Métricas qualitativas

Como apresentado na Subseção 2.6.2 essas métricas não necessitam de uma resposta de referencia esperada para serem geradas. Nessa etapa foi utilizado o *DeepEval*<sup>9</sup> para geração dessas métricas, a escolha do *DeepEval* em contraste com o *RAGAS* foi devido a vantagem deste em fornecer abstrações para a *LLM* critica utilizada, visto que no momento do experimentos o *RAGAS* tinha problemas para suporte de modelos da *Google*.

A seguir serão apresentando em detalhes como é feito essa avaliação.

##### Entradas

Após a etapa de estimular o sistema de dialogo com uma consultada de entrada. Conseguise isolar a saída de cada experimento. A Tabela 4.8 exemplifica as principais informações utilizadas para gerar as métricas qualitativas.

Nessa etapa também é utilizado uma *LLM* critica para fazer a avaliação. Convencionou-se utilizar o mesmo modelo de *LLM* utilizando durante a inferência para fazer a avaliação, ou seja, se a inferência foi feita utilizando o *Gemini-1.5-pro*, esse mesmo modelo será utilizado para a etapa de avaliação, com as mesmas configurações.

##### Saídas

Apos ser feita toda a etapa de avaliação, um conjunto de *scores* é apresentando para cada métrica junto com uma razão para da suporte ao *score* da métrica. Na Tabela 4.9 é apresentado um exemplo dessas saídas

Com essas avaliações feitas para todas as perguntas em todos os modelos é possível chegar em tendencias agrupadas por modelo, por pergunta, por paciente e etc.

---

<sup>9</sup><https://deepeval.com/>

Tabela 4.8: Resultado para a pergunta Q1 sobre medicações utilizando o modelo *Gemini-1.5-pro* utilizando a busca lexical sem expansão de grafos para a paciente *Beatris270\_Bogan287*

Consulta	Contexto	Resposta
What's my current medications and how should I be taking them?	{'Main health record': 'Medication Request Status: Stopped... (detalhes do registro médico) ...Reason: Streptococcal sore throat (disorder)'} }	You have two active medications:  <ul style="list-style-type: none"> <li>• <b>NDA020800 0.3 ML Epinephrine 1 MG/ML Auto-Injector:</b> This was prescribed by Dr. Alvin56 Crona259 on August 30, 2017. You should take this medication <b>as needed</b>.</li> <li>• <b>Fexofenadine hydrochloride 30 MG Oral Tablet:</b> This was prescribed by Dr. Alvin56 Crona259 on August 22, 2017. You should take this medication <b>as needed</b>.</li> </ul>

Tabela 4.9: Resultados dos scores qualitativos para as entradas da Tabela 4.8

Métrica	Score	Razão
HALS	0.0	<i>The score is 0.00 because the actual output aligns perfectly with the provided context, as indicated by the factual alignment statement. There are no contradictions present.</i>
ARS	1.0	<i>The score is 1.00 because the response effectively addresses the user's prompt, which asks for current medications and instructions on how to take them. Keep up the good work!</i>
CRS	0.333	<i>The score is 0.33 because some of the retrieved context is irrelevant. While the input asks about current medications and their dosage instructions, some of the retrieved context discusses medications that have been stopped, as indicated by the reasons for irrelevancy: "The status of the medication is 'Stopped' indicating it is no longer being taken." However, there are still relevant statements like "Medication: NDA020800 0.3 ML Epinephrine 1 MG/ML Auto-Injector, Dosage Instruction: Take as needed" and "Medication: Fexofenadine hydrochloride 30 MG Oral Tablet, Dosage Instruction: Take as needed", which slightly increases the relevancy score.</i>

#### 4.4.2 Métricas subjetivas

As métricas subjetivas foram feitas através de uma comparação pareada (*pairwise comparison*) entre as respostas de cada estratégia de busca. Ou seja, é feita avaliação em pares;

comparando a Saída da busca lexical sem expansão de grafos contra a saída da busca lexical com expansão de grafos, assim fazendo todas as comparações possíveis resultando em  $C(4, 2) = \frac{4!}{2!(4-2)!} = 6$  comparações.

## Entradas

Como apresentado na Tabela 4.10 as entradas dessa etapa são a consulta utilizada na fase de inferência, as respostas de cada estratégia de busca a ser comparadas. Nessa etapa utiliza-se o mesmo modelo de linguagem utilizado na etapa de inferência, com as configurações padrões. O Apêndice C mostra a instrução utilizada nessa etapa.

Tabela 4.10: Entradas para a avaliação subjetiva, respostas L0 e L1 utilizando o modelo *Gemini-1.5-pro* para a paciente *Beatris270\_Bogan287* após a consulta Q1

Resposta 1 - L1	Resposta 2 - L0
<p>You are currently taking two medications:</p> <ol style="list-style-type: none"> <li><b>NDA020800 0.3 ML Epinephrine 1 MG/ML Auto-Injector:</b> <ul style="list-style-type: none"> <li>Prescribed by Dr. Alvin56 Crona259 on August 30, 2017</li> <li>Take <b>as needed</b></li> <li>For: Allergy to substance</li> <li>Filled at HOLY FAMILY HOSPITAL</li> <li>Cost: \$64.66</li> </ul> </li> <li><b>Fexofenadine hydrochloride 30 MG Oral Tablet:</b> <ul style="list-style-type: none"> <li>Prescribed by Dr. Alvin56 Crona259 on August 22, 2017</li> <li>Take <b>as needed</b></li> <li>For: Perennial allergic rhinitis</li> <li>Filled at HOLY FAMILY HOSPITAL</li> <li>Cost: \$455.32</li> </ul> </li> </ol> <p><i>Note: Consult your doctor for updated information.</i></p>	<p>You have two active medications:</p> <ul style="list-style-type: none"> <li><b>NDA020800 0.3 ML Epinephrine 1 MG/ML Auto-Injector:</b> This was prescribed by Dr. Alvin56 Crona259 on August 30, 2017. You should take this medication <b>as needed</b>.</li> <li><b>Fexofenadine hydrochloride 30 MG Oral Tablet:</b> This was prescribed by Dr. Alvin56 Crona259 on August 22, 2017. You should take this medication <b>as needed</b>.</li> </ul>

## Saídas

O resultado dessa avaliação é um texto em estruturado em *JSON* com a escolha do modelo em cada métrica avaliada e a razão que suporta a escolha do modelo. Posteriormente é

feito um processamento nessa saída para extrair e apresentar os resultados. Importante ressaltar que nessa etapa o modelo pode não atribuir um vencedor claro: resultando em um “empate” na métrica avaliada. A Tabela 4.11 exemplifica esse processo de avaliação onde o modelo escolheu a resposta da saída do L1 para maior parte das métricas avaliadas nessa etapa.

Tabela 4.11: Avaliação subjetiva dada as entradas presentes na Tabela 4.10

Saida da avaliação subjetiva: L1 (Answer 1) contra L0 (Answer 2)
<p><b>Comprehensiveness: Winner: Answer 1.</b> Answer 1 provides more details about each medication, including the prescribing doctor, date prescribed, reason for prescription, pharmacy dispensed, cost, and date of related medical encounter. Answer 2 only includes the prescribing doctor, date prescribed, and dosage instructions.</p>
<p><b>Diversity: Winner: Answer 1.</b> Answer 1 offers a slightly more diverse range of information by including details about the prescription and medical encounter. While neither answer delves into different perspectives on the medications themselves, Answer 1’s additional details provide a broader context.</p>
<p><b>Empowerment: Winner: Answer 1.</b> Answer 1 empowers the reader with more information to understand their medication history and potentially discuss their prescriptions with their doctor or pharmacist. The additional details about cost and medical encounters could be valuable for patients.</p>
<p><b>Directness: Winner: Answer 2.</b> Answer 2 is slightly more direct in answering the question. It presents the essential information about the medications and dosage instructions concisely. Answer 1, while more comprehensive, includes additional details that might be considered extraneous to the core question.</p>
<p><b>Overall Winner: Winner: Answer 1.</b> While Answer 2 is more concise, Answer 1 is the overall winner due to its superior comprehensiveness and empowerment. The additional details provided in Answer 1 offer valuable context and information for the reader, potentially aiding in better understanding and management of their medications. However, it is crucial to emphasize that both answers appropriately advise consulting a doctor or pharmacist for personalized medical advice.</p>

### 4.4.3 Métricas objetivas

Para fazer uma análise mais holística, principalmente de entender o aumento de custo ao adicionar mais informação no contexto dos LLMs. Todas as etapas foram instrumentadas para aferir, principalmente, tempo de execução em cada etapa e a quantidade de *tokens* consumidos para fazer a inferência da consulta. Nessa instrumentação foi utilizado marcadores de tempo nativos da linguagem e para consumo de *tokens* foi utilizado ferramentas

nativas do *framework LangChain*<sup>10</sup>. A Tabela 4.12 ilustra algumas dessas métricas para a consulta Q1 utilizando o modelo *Gemini-1.5-pro* para a paciente *Beatris270\_Bogan287*

Tabela 4.12: Métricas de Tempo e *Tokens* por Tipo de Busca

<b>Tipo de Busca</b>	<b>Tempo Total (s)</b>	<b>Quantidade de <i>Tokens</i> Totais</b>
L0	8.899	1675.0
L1	10.544	4340.0
S0	5.471	2878.0
S1	18.407	6438.0

---

<sup>10</sup><https://www.langchain.com/>

# Capítulo 5

## Resultados

Este capítulo apresenta e discute os resultados obtidos a partir da metodologia proposta. Todos os códigos fontes e resultados podem ser encontrados no repositório publico do *GitHub*<sup>1</sup>

Os experimentos realizados concentraram-se na comparação entre os métodos de busca lexical e busca por similaridade, variando a quantidade de *k-hops* para expansão do grafo de 0 a 1. Para esses experimentos, foram utilizados os cinco pacientes da Tabela 4.1, aplicando as oito perguntas da Tabela 4.7 para os modelos *GPT-4o-2024-08-06*, *Gemini-1.5-pro*, *Claude-v3-opus*, *Llama-3-70b-instruct-awq* e *DeepSeek-R1-Distill-Llama-70B*, resultando em 800 interações (4 buscas x 5 pacientes x 8 perguntas x 5 modelo).

O restante deste capítulo é dividido em quatro seções: Na Seção 5.1 será apresentado a base de grafo gerada apos o pre-processamento proposto; na Seção 5.2 serão mostrados os resultados agregados dos experimentos; na Seção 5.3 é sumarizado as métricas avaliadas. Na Seção 5.4 apresenta-se resultados específicos e por fim uma breve discussão sobre os resultados na Seção 5.5.

### 5.1 Construção da base de grafos

Conforme descrito na seção 4.2, o processo de construção da base de grafos, que converte dados em formato *JSON* para uma representação gráfica, ilustrado na Figura 5.1. Nessa representação, cada nó corresponde a um recurso do modelo *FHIR*, enquanto os relacionamentos entre os recursos são representados por arestas. As categorias de recursos são diferenciadas visualmente por cores, para facilitar a visualização.

O fluxo completo desse processo é detalhado na Seção 4.2. A Figura 5.1 mostra um exemplo do grafo gerado para um dos pacientes, cada paciente tem seu próprio grafo de conhecimentos isolado.

---

<sup>1</sup><https://github.com/baptvit/graphrag-on-fhir>

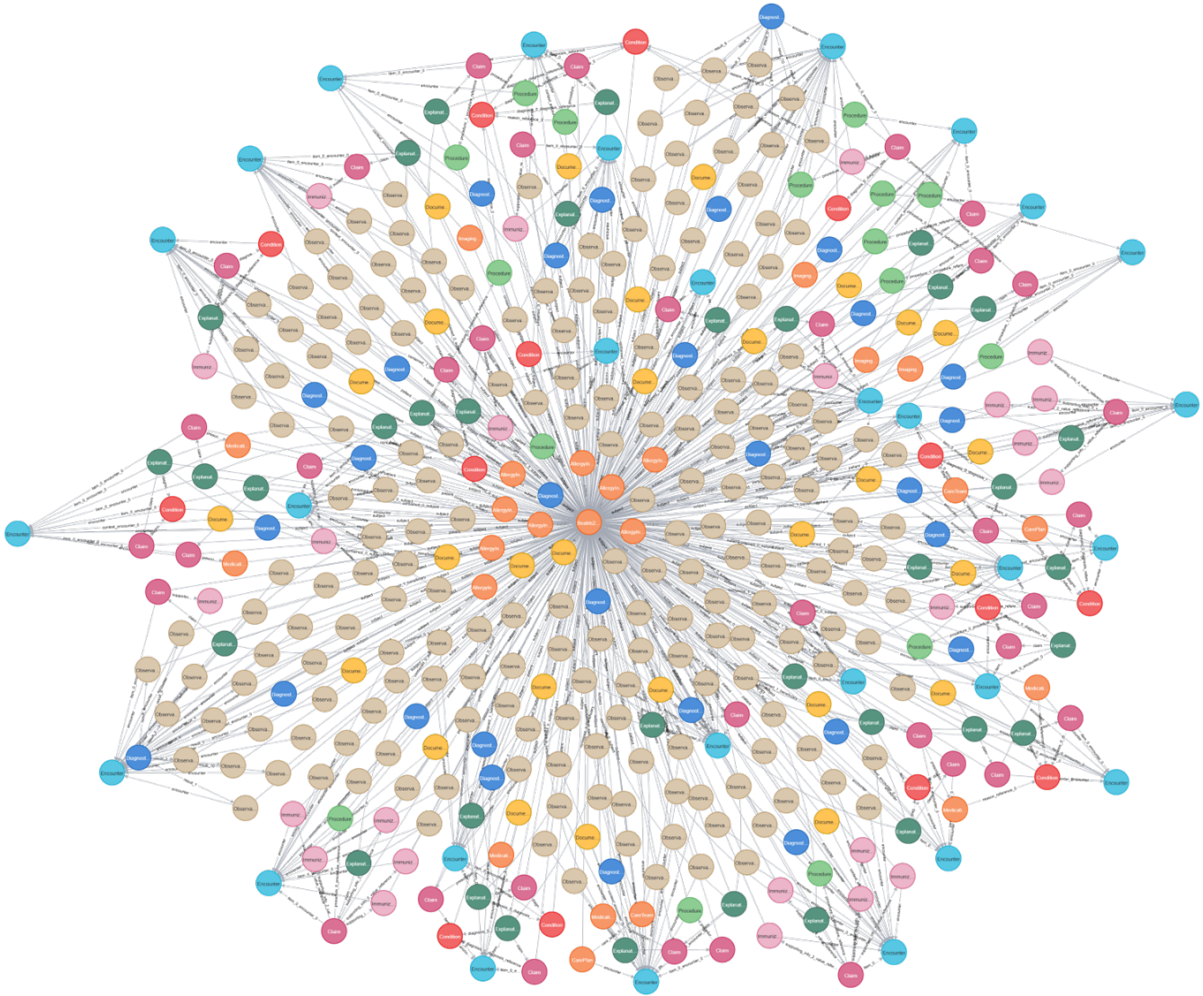


Figura 5.1: Representação gráfica da paciente *Beatris270 Bogan287* da Tabela 4.1.

Para os pacientes selecionados da Tabela 4.1, foi aplicado o pré-processamento dos recursos brutos para os recursos transformados como exemplificado na Tabela 4.4. Os resultados obtidos estão organizados conforme a tabela abaixo:

## 5.2 Experimentos

Com o intuito de validar a metodologia proposta apresentada neste trabalho, são analisadas as oito perguntas da Tabela 4.7 aplicada aos pacientes da Tabela 4.1, utilizando os modelos *GPT-4o-2024-08-06*, *Gemini-1.5-pro*, *Claude-v3-opus*, *Llama-3-70b-instruct-awq* e *DeepSeek-R1-Distill-Llama-70B* com um *system prompt* fixo (Apêndice A) para todos os experimentos.

As métricas avaliadas incluíram: *Context Relevancy*; *Answer Relevancy* e *Hallucination*. Essas métricas de avaliação qualitativas, como descrito na Seção 2.6, não necessita

Tabela 5.1: Resultado da redução de *Tokens* por paciente após o pré processamento

Paciente	Nº de Nós	Nº de Relacionamentos	Total de Tokens Brutos	Total de Tokens	Redução (%)
Beatris270 Bogan287	500	5.391	298.805	107.305	64,18%
Edythe31 Morar593	442	4.578	321.497	110.654	65,55%
Milton509 Ortiz186	1.129	6.088	567.063	214.344	62,19%
Allen322 Ferry57	1.235	6.532	803.945	276.071	65,65%
Jacklyn830 Veum823	2.405	13.818	1.534.203	558.554	63,59%
Média da Redução					64,23%

de uma referência de comparação esperada.

São replicados as métricas subjetivas dos trabalhos do Edge et al. (2024) e Guo et al. (2024), descritos na Seção 4.4, utilizando um *evaluation prompt* (Apêndice C) para comparações pares a pares entre os experimentos [18, 32].

### 5.3 Resultados gerais

Nesta seção, são apresentados os resultados quantitativos e qualitativos agregados, obtidos a partir da avaliação das diferentes estratégias de interação. Os dados são analisados tanto por questão quanto por paciente, focando-se na preferência proporcional e em métricas específicas de qualidade da resposta, em negrito estão apresentados os melhores resultados.

A Tabela 5.2 detalha a comparação do desempenho proporcional das estratégias L0, L1, S0 e S1 para cada uma das oito questões do estudo. A avaliação é segmentada por quatro critérios principais: *Comprehensiveness*, *Diversity*, *Empowerment* e *Directness*. Uma coluna adicional, *Overall Winner*, sintetiza a estratégia com melhor desempenho geral em cada questão. Observa-se que a estratégia L1 foi a vencedora geral na maioria das questões, especificamente em Q1, Q2, Q3, Q6, Q7 e Q8. A estratégia S1, por sua vez, destacou-se como vencedora nas questões Q4 e Q5. Analisando por critério, L1 também domina em *Comprehensiveness* na maioria dos casos. Nos demais critérios, há uma maior alternância entre as estratégias vencedoras.

De forma análoga, a Tabela 5.3 apresenta os mesmos dados de preferência proporcional, mas agregados por paciente. Ao analisar a coluna *Overall Winner*, verifica-se que a estratégia L1 foi a preferida para a maioria dos pacientes (Allen322, Beatris270, Edythe31 e Milton509). A estratégia S1 foi a vencedora para o paciente Jacklyn830. A estratégia L1

Tabela 5.2: Comparação de todas as estratégias por questão na escolha proporcional.

Métrica	Busca	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
<b>COMP</b>	L0	0.173	0.307	0.260	0.147	0.107	0.280	0.300	0.193
	L1	<b>0.327</b>	<b>0.394</b>	<b>0.447</b>	0.267	0.307	<b>0.327</b>	<b>0.427</b>	<b>0.353</b>
	S0	0.280	0.197	0.093	0.280	0.267	0.260	0.107	0.200
	S1	0.220	0.102	0.200	<b>0.307</b>	<b>0.320</b>	0.133	0.167	0.253
<b>DIV</b>	L0	0.140	0.312	0.233	0.160	0.113	<b>0.327</b>	0.320	0.213
	L1	<b>0.307</b>	<b>0.424</b>	<b>0.333</b>	0.267	0.253	0.300	<b>0.407</b>	0.280
	S0	0.247	0.184	0.113	0.267	0.280	0.167	0.093	0.160
	S1	<b>0.307</b>	0.080	0.320	<b>0.307</b>	<b>0.353</b>	0.207	0.180	<b>0.347</b>
<b>EMP</b>	L0	0.173	0.276	0.267	0.133	0.147	<b>0.273</b>	<b>0.340</b>	<b>0.313</b>
	L1	<b>0.307</b>	<b>0.362</b>	<b>0.347</b>	0.287	0.287	0.213	0.333	0.240
	S0	0.267	0.205	0.127	0.240	0.240	0.267	0.107	0.233
	S1	0.253	0.158	0.260	<b>0.340</b>	<b>0.327</b>	0.247	0.220	0.213
<b>DIR</b>	L0	<b>0.287</b>	0.312	<b>0.380</b>	0.107	<b>0.420</b>	0.247	<b>0.347</b>	<b>0.400</b>
	L1	0.260	<b>0.320</b>	0.260	0.349	0.280	0.233	0.173	0.145
	S0	0.213	0.216	0.193	0.168	0.153	0.187	0.260	0.269
	S1	0.240	0.152	0.167	<b>0.376</b>	0.147	<b>0.333</b>	0.220	0.186
<b>OW</b>	L0	0.193	0.291	0.287	0.133	0.127	0.273	0.327	0.240
	L1	<b>0.293</b>	<b>0.386</b>	<b>0.400</b>	0.287	0.273	<b>0.287</b>	<b>0.393</b>	<b>0.320</b>
	S0	0.240	0.205	0.087	0.247	0.273	0.247	0.087	0.187
	S1	0.273	0.118	0.227	<b>0.333</b>	<b>0.327</b>	0.193	0.193	0.253

mantém uma liderança consistente no critério *Comprehensiveness* para todos os pacientes. Para os critérios de *Diversity*, *Empowerment* e *Directness*, a liderança varia entre as quatro estratégias, dependendo do paciente.

A análise aprofunda-se nas métricas de quantitativas da busca, nomeadamente ARS (*Answer Relevance Score*), CRS (*Context Relevance Score*) e HALS (*Hallucination Score*). A Tabela 5.4 exhibe a pontuação média e o desvio padrão destas métricas para cada estratégia, por pergunta. Para a métrica ARS, as pontuações são consistentemente altas, frequentemente aproximando-se de 1.0, com destaque para as estratégias L1 e S0, que atingem a pontuação máxima em diversas questões. No que diz respeito à métrica CRS, a estratégia L1 tende a obter as maiores pontuações médias na maioria das questões. Para a métrica HALS, onde valores mais baixos são desejáveis, as estratégias L0 e S0 frequentemente apresentam as pontuações mais baixas, chegando a uma média de 0.0 em Q8.

Finalmente, a Tabela 5.5 fornece outra perspectiva, agregando as métricas de busca por paciente. Consistentemente com a análise por pergunta, a métrica ARS revela pontuações médias elevadas para todos os pacientes, com L1 e S0 apresentando frequentemente

Tabela 5.3: Comparação de todas as estratégias por paciente na escolha proporcional.

Métrica	Busca	Allen322	Beatris270	Edythe31	Jacklyn830	Milton509
<b>COMP</b>	L0	0.242	0.204	0.227	0.209	0.214
	L1	<b>0.338</b>	<b>0.375</b>	<b>0.380</b>	<b>0.312</b>	<b>0.371</b>
	S0	0.217	0.246	0.175	0.244	0.170
	S1	0.204	0.175	0.218	0.235	0.245
<b>DIV</b>	L0	0.242	0.221	0.233	0.197	0.236
	L1	<b>0.292</b>	<b>0.325</b>	<b>0.315</b>	<b>0.303</b>	<b>0.362</b>
	S0	0.183	0.242	0.164	0.222	0.131
	S1	0.283	0.213	0.289	0.278	0.271
<b>EMP</b>	L0	0.242	0.217	0.265	0.214	0.262
	L1	0.250	<b>0.317</b>	<b>0.299</b>	0.274	<b>0.341</b>
	S0	0.229	0.271	0.184	0.231	0.135
	S1	<b>0.279</b>	0.196	0.252	<b>0.282</b>	0.262
<b>DIR</b>	L0	<b>0.314</b>	<b>0.288</b>	<b>0.323</b>	<b>0.352</b>	<b>0.285</b>
	L1	0.255	0.283	0.237	0.209	0.272
	S0	0.213	0.188	0.190	0.235	0.211
	S1	0.218	0.242	0.250	0.204	0.233
<b>OW</b>	L0	0.250	0.221	0.244	0.222	0.227
	L1	<b>0.304</b>	<b>0.338</b>	<b>0.350</b>	<b>0.282</b>	<b>0.371</b>
	S0	0.200	0.254	0.150	0.231	0.144
	S1	0.246	0.188	0.256	0.265	0.258

Tabela 5.4: Comparação de estratégias de busca por pergunta

Métrica	Busca	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
<b>ARS</b>	L0	0.66±0.36	0.80±0.37	0.91±0.09	0.83±0.34	0.95±0.21	0.99±0.04	<b>0.95±0.21</b>	<b>0.99±0.04</b>
	L1	0.80±0.22	<b>0.83±0.34</b>	0.89±0.16	<b>0.99±0.04</b>	<b>0.99±0.02</b>	<b>1.00±0.01</b>	<b>0.95±0.08</b>	0.92±0.12
	S0	<b>0.91±0.17</b>	0.60±0.48	<b>0.92±0.21</b>	0.98±0.04	<b>0.99±0.03</b>	<b>1.00±0.02</b>	0.87±0.32	<b>0.99±0.04</b>
	S1	0.88±0.16	0.57±0.49	0.89±0.21	0.96±0.11	<b>0.99±0.04</b>	<b>1.00±0.00</b>	0.71±0.43	0.95±0.08
<b>CRS</b>	L0	0.40±0.22	0.43±0.48	0.67±0.22	0.62±0.26	<b>0.90±0.19</b>	0.64±0.18	0.84±0.19	<b>0.93±0.15</b>
	L1	0.47±0.25	<b>0.50±0.49</b>	<b>0.82±0.17</b>	0.62±0.18	0.76±0.26	<b>0.74±0.17</b>	<b>0.88±0.18</b>	0.88±0.16
	S0	<b>0.56±0.19</b>	0.36±0.39	0.55±0.30	0.62±0.25	0.52±0.26	0.68±0.16	0.82±0.25	0.90±0.11
	S1	0.52±0.20	0.37±0.43	0.73±0.22	<b>0.68±0.22</b>	0.60±0.23	0.63±0.24	0.71±0.17	0.81±0.18
<b>HALS</b>	L0	0.17±0.39	0.21±0.41	<b>0.05±0.20</b>	0.59±0.50	<b>0.04±0.21</b>	0.50±0.51	0.13±0.34	<b>0.00±0.00</b>
	L1	<b>0.15±0.37</b>	0.33±0.48	0.40±0.51	0.71±0.47	<b>0.04±0.21</b>	0.62±0.50	0.38±0.50	0.12±0.34
	S0	0.26±0.44	0.08±0.28	0.16±0.37	<b>0.26±0.45</b>	0.15±0.36	<b>0.40±0.50</b>	<b>0.12±0.33</b>	<b>0.00±0.00</b>
	S1	0.27±0.44	<b>0.06±0.22</b>	0.48±0.51	0.38±0.50	0.43±0.51	0.68±0.48	0.33±0.48	0.08±0.28

os melhores resultados. Para a métrica CRS, o desempenho varia, com L1, L0 e S0 liderando para diferentes pacientes. No que tange à métrica HALS, as estratégias L0 e S0 demonstram, em geral, as menores taxas de alucinação, registrando as médias mais baixas para a maioria dos pacientes.

Por fim, avalia-se a eficiência computacional das abordagens através das métricas de tempo de execução e consumo de tokens, fundamentais para a viabilidade prática do

Tabela 5.5: Comparação de estratégias de busca por Paciente

Estratégia	Nível	Allen322	Beatris270	Edythe31	Jacklyn830	Milton509
<b>ARS</b>	L0	0.93±0.18	0.88±0.30	0.87±0.27	<b>0.92±0.19</b>	0.82±0.35
	L1	<b>0.97±0.05</b>	<b>0.94±0.13</b>	0.88±0.21	<b>0.92±0.20</b>	0.87±0.23
	S0	0.96±0.16	0.88±0.28	<b>0.90±0.27</b>	0.89±0.29	<b>0.89±0.29</b>
	S1	0.87±0.31	0.88±0.26	0.88±0.27	0.86±0.33	0.83±0.30
<b>CRS</b>	L0	0.70±0.27	<b>0.85±0.20</b>	0.57±0.35	<b>0.64±0.32</b>	0.59±0.37
	L1	<b>0.79±0.23</b>	0.77±0.23	<b>0.65±0.34</b>	0.59±0.40	0.56±0.33
	S0	0.71±0.25	0.58±0.27	0.61±0.31	0.61±0.34	<b>0.63±0.34</b>
	S1	0.73±0.23	0.71±0.22	0.56±0.32	0.52±0.30	0.58±0.30
<b>HALS</b>	L0	<b>0.18±0.38</b>	0.10±0.31	0.22±0.42	0.29±0.46	0.32±0.48
	L1	0.29±0.46	0.20±0.40	0.40±0.50	0.27±0.46	0.45±0.51
	S0	0.27±0.44	<b>0.07±0.24</b>	<b>0.10±0.30</b>	0.42±0.50	<b>0.09±0.29</b>
	S1	0.65±0.48	0.17±0.38	0.26±0.44	<b>0.35±0.49</b>	0.25±0.43

sistema. A Tabela 5.6 sintetiza as estatísticas descritivas destas variáveis para as quatro estratégias. No que tange ao Tempo Total, as estratégias de busca sem expansão de grafos demonstram uma clara superioridade em desempenho. A estratégia S0 atinge a menor média de tempo e os menores quartis, seguida de perto pela L0, que apresenta a melhor mediana e o menor tempo máximo registrado. Em contrapartida, as estratégias que exploram a expansão por vizinhança L1 e S1 exigem um tempo de processamento sensivelmente maior, com médias superiores a 30 segundos.

Tabela 5.6: Estatísticas descritivas de Tempo e Tokens por Estratégia

Estatística	Tempo Total (s)			
	L0	L1	S0	S1
Média ± Std	23.57 ± 17.52	33.38 ± 26.76	<b>21.25 ± 15.49</b>	31.32 ± 25.93
Mínimo	4.06	4.09	<b>3.94</b>	4.15
25º Percentil	14.22	17.66	<b>12.41</b>	17.56
Mediana	<b>18.20</b>	25.88	18.37	26.15
75º Percentil	27.27	38.18	<b>25.13</b>	36.57
Máximo	<b>123.37</b>	171.89	127.70	218.72
Estatística	Total de Tokens			
	L0	L1	S0	S1
Média ± Std	<b>16,216 ± 22,641</b>	62,730 ± 70,435	17,544 ± 31,017	67,120 ± 94,814
Mínimo	<b>999</b>	1,005	1,034	1,032
25º Percentil	3,754	8,399	<b>3,553</b>	12,296
Mediana	<b>6,631</b>	35,252	6,644	40,299
75º Percentil	20,582	94,262	<b>17,594</b>	98,555
Máximo	<b>128,507</b>	395,997	279,085	786,994

Relativamente ao custo computacional medido pelo Total de Tokens, a disparidade entre as abordagens torna-se ainda mais evidente. A estratégia L0 confirma-se como a mais eficiente, registando consistentemente as menores valores em todas as estatísticas analisadas. Observa-se um aumento substancial no volume de tokens para as estratégias de L1 e S1, cujas médias ultrapassam os 60.000 tokens. A estratégia S1, em particular, apresenta o maior desvio padrão e o pico máximo absoluto, indicando que a expansão de um salto no grafo, embora potencialmente benéfica para o contexto, acarreta um custo operacional significativamente mais elevado. No Apêndice F encontra-se os resultados por modelo.

## 5.4 Resultados específicos

Nesta seção, aprofunda-se a análise da granularidade do desempenho do sistema, observando o comportamento das métricas quantitativas para cada paciente individualmente através das oito perguntas avaliadas. As Figuras 5.2, 5.3 e 5.4 apresentam os mapas de calor correspondentes às métricas ARS, CRS e  $HAL^{-1}$ , respectivamente.

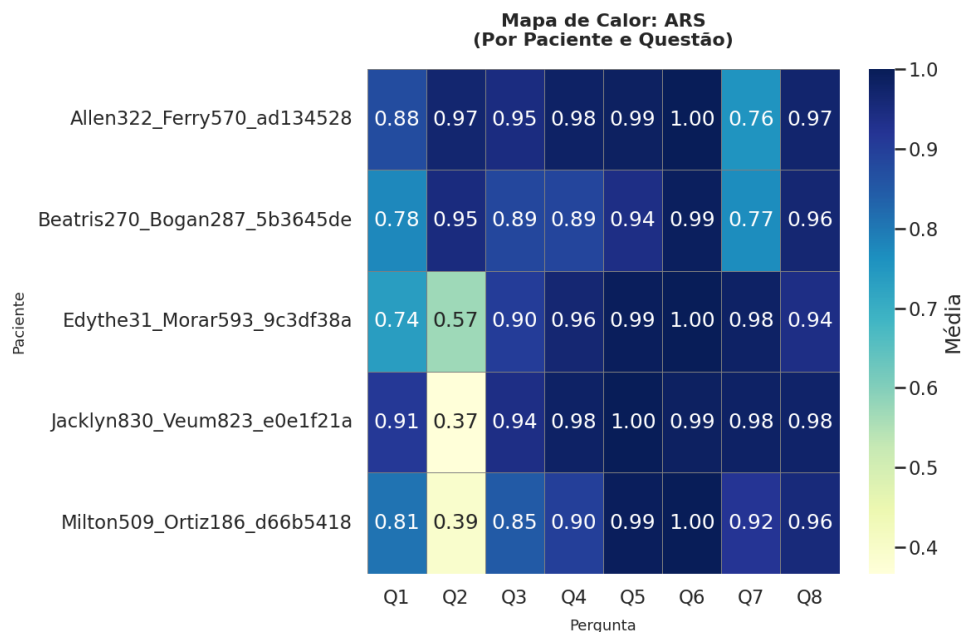


Figura 5.2: Mapa de calor das médias de relevância da resposta (ARS) por paciente e por pergunta.

A Figura 5.2 ilustra a ARS. Observa-se uma consistência notável no desempenho geral, com a maioria das interações resultando em pontuações superiores a 0,90. As questões Q5, Q6 e Q8 demonstraram ser as mais robustas, com o paciente *Allen322* atingindo a pontuação máxima na Q6 e *Jacklyn830* e *Edythe31* aproximando-se da perfeição na Q5.

No entanto, a Questão Q2 revelou-se um ponto crítico de falha para pacientes específicos. Enquanto *Allen322* e *Beatris270* mantiveram pontuações altas, os pacientes *Jacklyn830* e *Milton509* sofreram quedas abruptas de desempenho, registrando 0,36 e 0,39, respectivamente. Essa discrepância sugere que a complexidade ou a especificidade da Q2 afeta a capacidade do modelo de gerar respostas relevantes dependendo do histórico clínico do paciente.

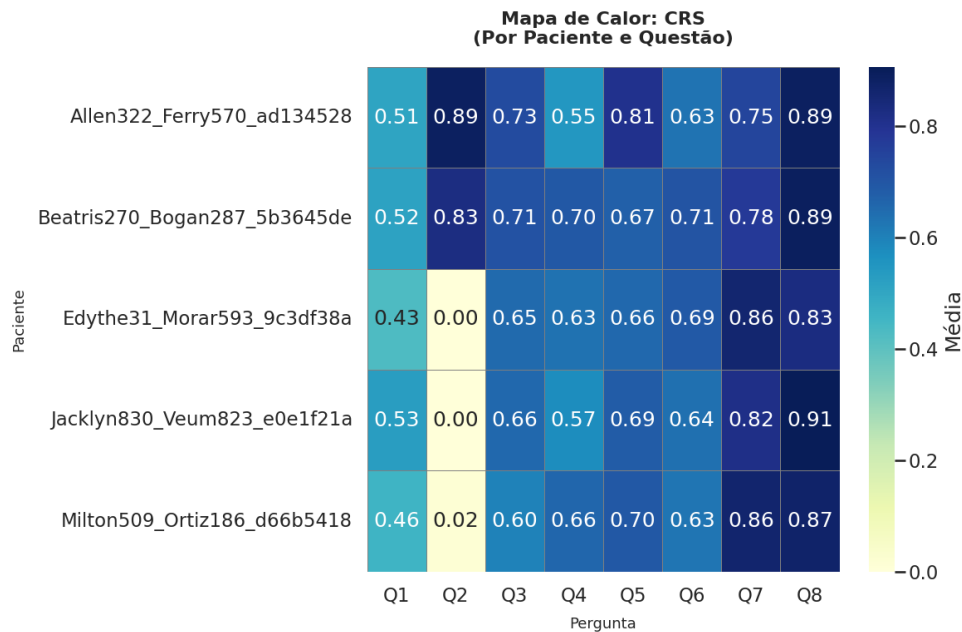


Figura 5.3: Mapa de calor das médias de relevância do contexto (CRS) por paciente e pergunta.

A análise da CRS, apresentada na Figura 5.3, expõe a maior variabilidade do sistema de recuperação. Diferente da ARS, onde as falhas foram pontuais, a recuperação de contexto apresentou desafios sistêmicos na Questão Q2. Os pacientes *Edythe31* e *Jacklyn830* obtiveram pontuação nula nesta questão, e *Milton509* obteve um valor residual de 0,01, indicando uma falha total ou falta de informação relevante para ser recuperada dos prontuários destes pacientes. Curiosamente, para o paciente *Allen322*, o sistema foi capaz de recuperar o contexto adequadamente na mesma questão, evidenciando a heterogeneidade dos dados de saúde.

Por outro lado, a Questão Q8 mostrou-se consistentemente para o recuperador, com pontuações entre 0,83 e 0,90 para todos os pacientes, sugerindo que a terminologia ou a estrutura da informação solicitada nesta questão está bem alinhada com a indexação dos documentos.

A métrica de Fidelidade ( $HAL^{-1}$ ), onde valores próximos a 1,0 indicam menor indicativo de alucinação, é detalhada na Figura 5.4. O comportamento aqui é distinto

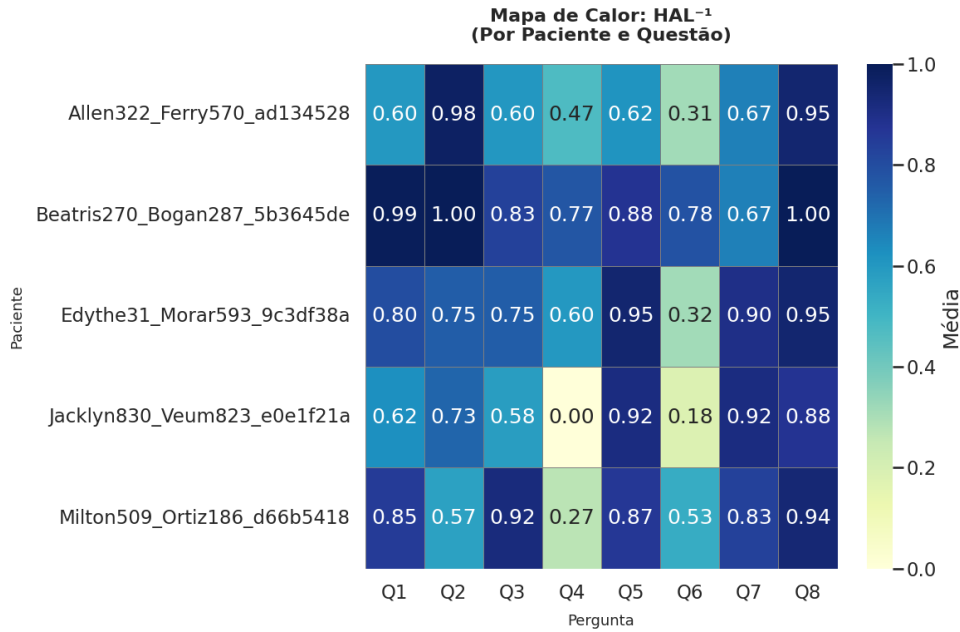


Figura 5.4: Mapa de calor das médias de Fidelidade  $HAL^{-1}$  por paciente por pergunta.

das métricas anteriores. O paciente *Beatris270* apresentou o desempenho mais seguro e estável, com fidelidade perfeita nas questões Q2 e Q8, e nunca baixando de 0,66.

Entretanto, identificaram-se zonas de alto risco de alucinação. A Questão Q4 foi problemática para a paciente *Jacklyn830*, resultando em uma fidelidade de 0.0 alucinação total ou resposta totalmente fora do esperado, e problemática para *Milton509*. Similarmente, a Questão Q6, apesar de ter alta ARS, apresentou baixos índices de fidelidade para *Jacklyn830*, *Allen322* e *Edythe31*. Isso indica um cenário perigoso onde o modelo fornece uma resposta que parece muito relevante, mas que possivelmente não está fundamentada no texto recuperado  $HAL^{-1}$  baixa.

Em suma, enquanto a Questão Q8 se consolida como o melhor cenário, alta relevância, alto contexto e alta fidelidade para quase todos, as Questões Q2, Q4 e Q6 exigem atenção redobrada, pois expõem vulnerabilidades específicas dependendo do perfil do paciente carecendo de maior atenção em uma futura análise com especialistas em saúde.

## 5.5 Discussão

Esta seção analisa os resultados sob duas perspectivas fundamentais: o desempenho comparativo das estratégias de busca e o impacto das características dos dados clínicos na qualidade da resposta gerada.

### 5.5.1 Desempenho das Estratégias

A avaliação geral aponta uma preponderância da estratégia L1, que se consagrou vencedora na maioria das questões Q1, Q2, Q3, Q6, Q7 e Q8. Este resultado sugere que, no domínio médico, a exatidão terminológica é prioritária. Entidades como nomes de medicamentos como *Fexofenadine* ou diagnósticos específicos são recuperados com maior precisão via correspondência exata da tipologia dos recursos FHIR do que por aproximação vetorial para os dados do presente trabalho com o modelo vetorial utilizado.

No entanto, a estratégia S1 demonstrou ser superior em questões conceituais e interpretativas, como Q4 que aborda Valores laboratoriais e Q5 que aborda Plano de cuidados. Nestes casos, onde a pergunta exige uma síntese “*summarize*”, “*what do they mean*” em vez de uma extração factual simples, a busca semântica capturou melhor a intenção do usuário. Este resultado corrobora os achados de Edgar et al. (2024) [18], que demonstram que abordagens baseadas em grafos superam a recuperação tradicional em tarefas de “resumo focado em consultas” (*query-focused summarization*), onde a compreensão de relacionamentos globais entre entidades é mais crítica do que a recuperação de um fato isolado.

Dois fenômenos críticos emergem desta comparação:

1. **O Custo do Contexto:** As estratégias de expansão com 1 salto triplicam o consumo de *tokens* e aumentam o tempo de resposta em cerca de 50% visto pela Tabela 5.6. Embora melhorem a relevância da resposta, elas inviabilizam aplicações sensíveis à latência ou custo.
2. **O Paradoxo da Fidelidade:** Mais contexto não garante maior factualidade. As estratégias L0 e S0 apresentaram, em média, menores taxas de possíveis alucinações do que as estratégias expandidas com grafos. A inclusão de nós vizinhos no grafo parece introduzir ruído informativo. Este comportamento alinha-se ao fenômeno descrito por Liu et al. (2023) [25], onde modelos de linguagem demonstram degradação de desempenho quando informações relevantes são “diluídas” em contextos longos, dificultando a distinção entre sinal e ruído.

### 5.5.2 O Impacto do Perfil Clínico na Recuperação

A análise granular por paciente revela que “não existe uma solução única para todos os casos”. A eficácia do *RAG* varia drasticamente conforme a densidade, a natureza dos dados presentes no prontuário FHIR e ao tipo de pergunta utilizada.

Pacientes com histórico clínico massivo impõem um “teste de estresse” ao sistema. A paciente *Jacklyn830*, com o maior volume de recursos 2.405, sendo 1.488 observações,

sofreu com alucinações severas na Q4 relacionadas a Laboratório, registrando fidelidade nula em algumas execuções. O excesso de dados históricos provavelmente saturou a capacidade do recuperador de distinguir o que era “recente”, levando à possível fabricação de valores.

Este cenário exemplifica empiricamente o desafio apontado por Liu et al. (2024) [25], onde o modelo falha em acessar informações precisas quando o contexto de entrada é excessivamente longo ou quando a resposta alvo não está nas extremidades da instrução de contexto. Para pacientes complexos como *Jacklyn830*, a expansão de contexto por grafos exacerbou esse problema, tornando a recuperação de fatos específicos menos confiável do que em pacientes com históricos concisos, como a paciente pediátrica *Beatris270*.

Outra análise derivada da visão paciente-pergunta, o sistema demonstrou incapacidade de lidar com a “ausência de evidência”. Na questão Q2 sobre alergias:

- Para pacientes com alergias explícitas como: *Beatris270*, a recuperação foi perfeita.
- Para pacientes sem alergias registradas como: *Milton509*, *Jacklyn830*, o sistema falhou gerando CRS próximo a 0.

O sistema busca documentos que contenham a palavra “alergia” ou semântica próxima; na ausência destes, ele não consegue inferir a negativa de “Não constam alergias”, resultando em respostas alucinadas ou vazias de contexto. Isso pode ter sido tanto na etapa de geração quanto na etapa de validação quantitativas das respostas onde o *framework* de avaliação não conseguiu lidar com esse caso de forma propriamente eficaz.

Na questão Q6 sobre valores financeiros, observou-se o fenômeno da “alucinações intrínsecas” (*intrinsic hallucinations*) referem-se às saídas do modelo que conflitam diretamente com o contexto de origem fornecido [51]. O modelo gerou respostas com alta Relevância ARS  $\approx 1.0$ , indicando estar formatadas corretamente como contas médicas, mas com alta probabilidade de alucinação. O *LLM* possivelmente gerou valores que pareciam corretos para o avaliador, mas que não correspondiam aos dados factuais dos recursos de *Claim*, possivelmente evidenciando a limitação dos modelos de linguagem em tarefas aritméticas precisas sem o auxílio de ferramentas externas como apresentado por Boye et al. (2025) [52] isso se torna mais latente para os modelos de linguagem utilizados na etapa de inferência do trabalho.

# Capítulo 6

## Conclusões

Este trabalho debruçou-se sobre um desafio central na saúde digital: a dificuldade de pacientes e profissionais em democratizar, acessar e compreender a vasta quantidade de informações contidas em Prontuários Eletrônicos de Pacientes, especialmente quando estruturados no complexo padrão FHIR. A investigação partiu da premissa de que a abordagem de *GraphRAG* poderia superar as limitações de sistemas de *RAG* tradicionais, que frequentemente falham em capturar a natureza interconectada dos dados clínicos. O objetivo principal foi, portanto, validar a eficácia da *GraphRAG* na interpretação de dados FHIR, avaliando se a exploração explícita das relações em grafos, em grafos de conhecimentos, resultaria em respostas de maior qualidade, relevância e detalhe.

Os resultados obtidos confirmaram a hipótese principal de que a *GraphRAG*, implementada através da expansão de vizinhos de 1 salto, demonstrando melhora significativa nas métricas subjetivas de avaliação em comparação com as abordagens de *RAG* tradicionais. A estratégia L1 destacou-se na maioria das questões que exigem precisão factual e terminológica, como a recuperação de medicamentos da questão Q1 e condições médicas da questão Q3, sendo eleita como a vencedora geral em seis das oito perguntas avaliadas. Por outro lado, a estratégia S1 mostrou-se mais eficaz em questões conceituais que demandavam síntese e interpretação, como o resumo de valores laboratoriais da questão Q4, corroborando a hipótese secundária de que diferentes tipos de busca se adequam a diferentes naturezas de consulta.

Apesar do sucesso em melhorar a qualidade das respostas, a investigação revelou um paradoxo fundamental: mais contexto nem sempre equivale a maior qualidade e fidelidade. A expansão do grafo, embora enriquecedora, aumentou o risco de alucinações, especialmente em prontuários com grande volume de dados. Este fenômeno, alinhado ao problema de “contexto perdido no meio” (*lost in the middle*), demonstrou que a inclusão de um volume massivo de informações pode diluir a relevância dos fatos cruciais, levando o *LLM* a gerar informações plausíveis, mas não factualmente ancoradas no contexto re-

cuperado. Ademais, o ganho em qualidade veio a um custo computacional considerável, com um aumento expressivo na latência e no consumo de *tokens*, um fator crítico para a viabilidade de implementação em sistemas produtivos.

Conclui-se, portanto, que a *GraphRAG* é uma abordagem promissora e poderosa para a interpretação de dados FHIR, capaz de gerar respostas mais completas e contextualmente ricas e detalhadas, sendo uma técnica que aproveita a semântica do modelo FHIR de forma efetiva. No entanto, sua aplicação bem-sucedida não é trivial e exige um balanço cuidadoso entre a riqueza do contexto recuperado, o custo computacional e o risco de degradação da atenção do modelo. A eficácia do sistema mostrou-se altamente dependente do perfil clínico do paciente e da natureza da pergunta, evidenciando que uma solução única para todos os cenários é inadequada para a complexidade do domínio da saúde.

## 6.1 Contribuições do Trabalho

As contribuições desta dissertação para a área de saúde digital e inteligência artificial são diversas e inter-relacionadas. A principal delas é a validação empírica da abordagem *GraphRAG* em dados FHIR, demonstrando que a expansão de contexto via grafos resulta em respostas qualitativamente superiores. Este avanço foi medido através de uma análise comparativa abrangente entre estratégias de recuperação lexical e por similaridade, o que permitiu identificar as forças e fraquezas de cada método e oferecer um guia prático sobre qual abordagem adotar com base na natureza da consulta. Essa análise foi possibilitada pela proposição de uma metodologia de avaliação holística, que combinou métricas qualitativas, subjetivas via *LLM-as-a-judge* e métricas objetivas de desempenho. Este *framework* de avaliação multifacetado não apenas validou a abordagem, mas também permitiu a identificação de modos de falha críticos em sistemas *RAG* aplicados à saúde, como o “paradoxo da fidelidade”, a incapacidade de lidar com a ausência de evidência e a tendência a alucinações plausíveis. Como resultado direto da pesquisa, destaca-se também a produção bibliográfica com a publicação do artigo “*GraphRAG on Electronic Health Record: A Knowledge Graph-Enhanced RAG Approach for Healthcare Information Access*”[19], disseminando o conhecimento gerado na comunidade científica internacional.

## 6.2 Limitações do Estudo

Apesar dos resultados promissores, este estudo possui limitações que devem ser consideradas e que abrem caminhos para futuras pesquisas. Primeiramente, a pesquisa foi conduzida utilizando um conjunto de dados sintéticos e a avaliação foi realizada por um *LLM* juiz, em vez de prontuários médicos reais e da validação por especialistas clíni-

cos. Essa escolha, embora escalável, pode limitar a generalização dos resultados para a complexidade, o ruído e a heterogeneidade dos dados do mundo real, e não substitui o padrão-ouro da validação clínica que envolve necessariamente avaliação por especialistas. Adicionalmente, o escopo técnico da investigação foi contido, explorando apenas uma expansão de 1 salto e utilizando uma abordagem de truncamento simples para o gerenciamento de contexto, sem implementar técnicas mais sofisticadas para mitigar o problema de contexto perdido no meio [25]. Por fim, a metodologia revelou uma vulnerabilidade notável na incapacidade do sistema em tratar adequadamente a “ausência de evidência”, um cenário comum e crítico em aplicações clínicas onde o sistema falha em informar que um dado não existe no prontuário.

### 6.3 Trabalhos Futuros

Com base nas conclusões e limitações identificadas, diversas direções para trabalhos futuros emergem. O passo mais crucial é estender a validação da arquitetura *GraphRAG* para conjuntos de dados de pacientes reais e anonimizados, em maior escala, e conduzir estudos de avaliação com especialistas da área da saúde para aferir a acurácia clínica e a segurança das respostas. Paralelamente, é imperativo o desenvolvimento de estratégias avançadas de gerenciamento de contexto para mitigar o fenômeno de “contexto perdido no meio”, como o re-ranqueamento de nós, a sumarização inteligente de subgrafos e o uso de *prompting* estruturado. Sugere-se também a implementação de mecanismos de recuperação híbridos e adaptativos, que selecionem dinamicamente a estratégia de busca mais apropriada com base na intenção da pergunta. Adicionalmente, futuras iterações do sistema devem aprimorar o tratamento da ausência de evidência, capacitando o modelo a reconhecer e comunicar quando uma informação não está disponível. Finalmente, a arquitetura pode ser expandida para incorporar o uso de ferramentas externas para tarefas específicas, como cálculos precisos, e para enriquecer o grafo de conhecimento conectando-o a bases de conhecimento médico externas e autoritativas, a fim de aumentar a factualidade e prover respostas verdadeiramente baseadas em evidências.

# Referências

- [1] Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang e Haofen Wang: *Retrieval-augmented generation for large language models: A survey*. ArXiv, abs/2312.10997, 2023. <https://api.semanticscholar.org/CorpusID:266359151>. x, 7, 8, 9, 14, 30
- [2] Peng, Boci, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang e Siliang Tang: *Graph retrieval-augmented generation: A survey*, 2024. <https://arxiv.org/abs/2408.08921>. x, 2, 3, 10, 11, 12, 31
- [3] Schmiedmayer, Paul, Adrit Rao, Philipp Zagar, Vishnu Ravi, Aydin Zahedivash, Arash Fereydooni e Oliver Aalami: *Llm on fhir – demystifying health records*, 2024. <https://arxiv.org/abs/2402.01711>. xi, 1, 2, 3, 24, 25, 26, 27, 28, 29, 36, 37, 39, 40
- [4] Guo, Jiayan, Lun Du, Hengyu Liu, Mengyu Zhou, Xinyi He e Shi Han: *Gpt4graph: Can large language models understand graph structured data ? an empirical evaluation and benchmarking*, 2023. <https://arxiv.org/abs/2305.15066>. xi, 12, 23, 37, 38
- [5] *Criteria for selecting the leading health indicators for healthy people 2030*, agosto 2019. <https://www.ncbi.nlm.nih.gov/books/NBK552637/>. 1, 2
- [6] Bender, Duane e Kamran Sartipi: *HL7 fhir: An agile and restful approach to healthcare information exchange*. Em *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, páginas 326–331, 2013. 1, 2, 20, 30
- [7] 21st Century Cures Act: *21st century cures act: An act to accelerate the discovery, development, and delivery of 21st century cures, and for other purposes*, december 2016. <https://www.govinfo.gov/app/details/PLAW-114publ255>. 1
- [8] Resources, HRSA Health e Services Administration: *Addressing health literacy*, jun 2024. <https://www.hrsa.gov/about/organization/bureaus/ohe/health-literacy>. 1, 2
- [9] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser e Illia Polosukhin: *Attention is all you need*. Em Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan e R. Garnett (editores): *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf). 1, 7

- [10] Schifman, Sam: *Rag on fhir with knowledge graphs*, jan 2024. <https://medium.com/@samschifman/rag-on-fhir-with-knowledge-graphs-04d8e13ee96e>. 1, 2, 3, 23, 27, 31, 32, 36
- [11] Labs, KMS Healthcare Innovation: *Using language models for extracting insights from patient's medical records: A research & development perspective*. Tech report. <https://kms-healthcare.com/wp-content/uploads/2024/08/Research-Paper-Using-LLM-for-Extracting-Insights-from-Patients-Medical-Records.pdf>. 1, 3, 24, 26, 27, 31, 36, 37, 39
- [12] Ge, J., S. Sun, J. Owens, V. Galvez, O. Gologorskaya, J. C. Lai, M. J. Pletcher e K. Lai: *Development of a liver disease-specific large language model chat interface using retrieval augmented generation*. medRxiv : the preprint server for health sciences, 2023.11.10:23298364, 2023. <https://doi.org/10.1101/2023.11.10.23298364>. 1, 22
- [13] Van Veen, D., C. Van Uden e L. et al. Blankemeier: *Adapted large language models can outperform medical experts in clinical text summarization*. Nature Medicine, 30(11):1134–1142, 2024. <https://doi.org/10.1038/s41591-024-02855-5>. 1, 3, 22, 25, 40
- [14] Zaretsky, Jonah, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B. Blecker e Jonah Feldman: *Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format*. JAMA Network Open, 7(3):e240357–e240357, março 2024, ISSN 2574-3805. <https://doi.org/10.1001/jamanetworkopen.2024.0357>. 1, 22, 25, 26
- [15] Saba, Walid, Suzanne Wendelken e James. Shanahan: *Question-answering based summarization of electronic health records using retrieval augmented generation*. ArXiv, abs/2401.01469, 2024. <https://api.semanticscholar.org/CorpusID:266741989>. 1, 22, 25, 37
- [16] Wu, Junde, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina e Vicente Grau: *Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation*, 2024. <https://arxiv.org/abs/2408.04187>. 1, 2, 3, 22, 25
- [17] Elgedawy, Ran, Ioana Danciu, Maria Mahbub e Sudarshan Srinivasan: *Dynamic q&a of clinical documents with large language models*, 2024. <https://arxiv.org/abs/2401.10733>. 1, 22, 37
- [18] Edge, Darren, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt e Jonathan Larson: *From local to global: A graph rag approach to query-focused summarization*, 2024. <https://arxiv.org/abs/2404.16130>. 4, 12, 20, 23, 26, 27, 39, 40, 48, 55

- [19] Baptista, João V. R. e Luís P. F. Garcia: *Graphrag on electronic health record: A knowledge graph-enhanced rag approach for healthcare information access*. Em Freitas, Rosiane de e Diego Furtado (editores): *Intelligent Systems*, páginas 615–630, Cham, 2026. Springer Nature Switzerland, ISBN 978-3-032-15987-8. 4, 58
- [20] Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian Yun Nie e Ji Rong Wen: *A survey of large language models*, 2025. <https://arxiv.org/abs/2303.18223>. 7
- [21] OpenAI: *Gpt-4 technical report*, 2024. <https://arxiv.org/abs/2303.08774>. 7
- [22] Li, Tianle, Ge Zhang, Quy Duc Do, Xiang Yue e Wenhua Chen: *Long-context llms struggle with long in-context learning*, 2024. <https://arxiv.org/abs/2404.02060>. 10
- [23] Yang, Chia Jeng: *Graph vs. vector rag – benchmarking, optimization levers, and a financial analysis example*. <https://medium.com/neo4j/graph-vs-vector-rag-benchmarking-optimization-levers-and-a-financial-analysis-exa>. 10
- [24] Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov e Christopher D. Manning: *HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering*. Em *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 2871–2880, Brussels, Belgium, 2018. Association for Computational Linguistics. 14
- [25] Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Cheung, Percy S. Liang e Matthew S. Levine: *Lost in the Middle: How Language Models Use Long Contexts*. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. 15, 55, 56, 59
- [26] Manning, Christopher D., Prabhakar Raghavan e Hinrich Schütze: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008, ISBN 978-0-521-86571-5. 16
- [27] Järvelin, Kalervo e Jaana Kekäläinen: *Cumulated gain-based evaluation of IR techniques*. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. 17
- [28] Papineni, Kishore, Salim Roukos, Todd Ward e Wei Jing Zhu: *BLEU: a Method for Automatic Evaluation of Machine Translation*. Em *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 311–318, Philadelphia, Pennsylvania, 2002. Association for Computational Linguistics. 18
- [29] Lin, Chin Yew: *ROUGE: A Package for Automatic Evaluation of Summaries*. Em *Proceedings of the Workshop on Text Summarization Branches Out (WAS)*, páginas 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 18
- [30] Hallucination: *Deepeval documentation*, dec 2024. <https://docs.confident-ai.com/docs/metrics-hallucination>. 20

- [31] Zheng, Lianmin, Wei Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez e Ion Stoica: *Judging llm-as-a-judge with mt-bench and chatbot arena*, 2023. <https://arxiv.org/abs/2306.05685>. 20
- [32] Guo, Zirui, Lianghao Xia, Yanhua Yu, Tu Ao e Chao Huang: *Lightrag: Simple and fast retrieval-augmented generation*, 2024. <https://arxiv.org/abs/2410.05779>. 20, 40, 48
- [33] Mandel, J. C., D. A. Kreda, K. D. Mandl, I. S. Kohane e R. B. Ramoni: *Smart on fhir: a standards-based, interoperable apps platform for electronic health records*. Journal of the American Medical Informatics Association : JAMIA, 23(5):899–908, 2016. 21
- [34] Sharma, Meenakshi e Himanshu Aggarwal: *Hl-7 based middleware standard for healthcare information system: Fhir*. Proceedings of 2nd International Conference on Communication, Computing and Networking, 2018. <https://api.semanticscholar.org/CorpusID:70156753>. 21
- [35] Es, Shahul, Jithin James, Luis Espinosa-Anke e Steven Schockaert: *Ragas: Automated evaluation of retrieval augmented generation*, 2023. <https://arxiv.org/abs/2309.15217>. 23, 27, 40
- [36] Ip, Jeffrey: *The open-source llm evaluation framework*. DeepEval Documentation, dec 2024. <https://docs.confident-ai.com/>. 23, 27, 40
- [37] Guo, Tiezheng, Qingwen Yang, Chen Wang, Yanyi Liu, Pan Li, Jiawei Tang, Dapeng Li e Yingyou Wen: *Knowledgenavigator: leveraging large language models for enhanced reasoning over knowledge graph*. *Complex & Intelligent Systems*, 10(5):7063–7076, julho 2024, ISSN 2198-6053. <http://dx.doi.org/10.1007/s40747-024-01527-8>. 23, 25, 37
- [38] Cheng, Keyuan, Gang Lin, Haoyang Fei, Yuxuan zhai, Lu Yu, Muhammad Asif Ali, Lijie Hu e Di Wang: *Multi-hop question answering under temporal knowledge editing*, 2024. <https://arxiv.org/abs/2404.00492>. 23, 37
- [39] Hu, Yuntong, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling e Liang Zhao: *Grag: Graph retrieval-augmented generation*, 2024. <https://arxiv.org/abs/2405.16506>. 23, 37
- [40] Gao, Hanning, Lingfei Wu, Po Hu, Zhihua Wei, Fangli Xu e Bo Long: *Graph-augmented learning to rank for querying large-scale knowledge graph*, 2022. <https://arxiv.org/abs/2111.10541>. 23, 25, 37
- [41] Baek, Jinheon, Alham Fikri Aji e Amir Saffari: *Knowledge-augmented language model prompting for zero-shot knowledge graph question answering*, 2023. <https://arxiv.org/abs/2306.04136>. 23, 37
- [42] Choudhary, Nurendra e Chandan K. Reddy: *Complex logical reasoning over knowledge graphs using large language models*, 2024. <https://arxiv.org/abs/2305.01157>. 23, 37

- [43] Dehghan, Mohammad, Mohammad Ali Alomrani, Sunyam Bagga, David Alfonso-Hermelo, Khalil Bibi, Abbas Ghaddar, Yingxue Zhang, Xiaoguang Li, Jianye Hao, Qun Liu, Jimmy Lin, Boxing Chen, Prasanna Parthasarathi, Mahdi Biparva e Mehdi Rezagholizadeh: Ewek-qa: Enhanced web and efficient knowledge graph retrieval for citation-based question answering systems, 2024. <https://arxiv.org/abs/2406.10393>. 23
- [44] Johnson, Alistair E. W., Tom J. Pollard, Lu Shen et al.: Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. <https://doi.org/10.1038/sdata.2016.35>. 24
- [45] Johnson, Alistair E. W., Lucas Bulgarelli, Lu Shen et al.: Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. <https://doi.org/10.1038/s41597-022-01899-x>. 24
- [46] Waisbord, Silvio: Fake health news in the new regime of truth and (mis)information. *RECIIS: Revista Eletrônica de Comunicação, Informação e Inovação em Saúde*, 14(1), 2020. <https://doi.org/10.29397/reciis.v14i1.1953>. 24
- [47] Kotonya, Neema e Francesca Toni: Explainable automated fact-checking for public health claims. *Em Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 7740–7754, Online, novembro 2020. Association for Computational Linguistics. <https://www.aclweb.org/anthology/2020.emnlp-main.623>. 24
- [48] Wu, Chaoyi, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang e Weidi Xie: Pmc-llama: Towards building open-source language models for medicine, 2023. <https://arxiv.org/abs/2304.14454>. 24
- [49] National Library of Medicine (US): Umls knowledge sources, May 2024. <http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html>, acesso em 2024-07-15, Dataset on the Internet. 24
- [50] He, Jia, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang e Sadid Hasan: Does prompt formatting have any impact on llm performance?, 2024. <https://arxiv.org/abs/2411.10541>. 31
- [51] Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin e Ting Liu: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, janeiro 2025, ISSN 1558-2868. <http://dx.doi.org/10.1145/3703155>. 56
- [52] Boye, Johan e Birger Moell: Large language models and mathematical reasoning failures, 2025. <https://arxiv.org/abs/2502.11574>. 56

# Apêndice A

## Instrução de sistema

Tabela A.1: Instrução de sistema para todos os experimentos

<b>Instrução de sistema utilizado nos experimentos</b>
<p>You are a dedicated and highly skilled FHIR Specialist. Your task is to provide in-depth, comprehensive responses to patient queries, focusing on clarity, accuracy, and relevance. In each response, prioritize the inclusion of all key details available within the patient's record, such as:</p> <ul style="list-style-type: none"><li>- Dates of service, appointments, or medical events</li><li>- Healthcare providers involved, including doctors, specialists, or hospitals</li><li>- Locations of care, specific departments, or facilities</li><li>- Billing information, such as charges, payment status, and insurance details</li><li>- Other relevant data that may be beneficial for patient understanding and decision-making</li></ul> <p>Consider the broader implications of your response to ensure it aids the patient's understanding and supports informed decision-making. Carefully review the entire context before answering to ensure your response is exhaustive and precisely tailored. Aim for depth and clarity in your explanations, ensuring that your response addresses the patient's needs and potential concerns fully and thoughtfully.</p>

# Apêndice B

## Instrução de Pre-processamento

Tabela B.1: Instrução de pré-processamento do recurso FHIR em formato bruto para o formato de texto

<b>Instrução de pré processamento dos recursos FHIR brutos</b>
As an expert in FHIR R4 conversion, please convert the following FHIR R4 resource into a concise, human-readable text. Ensure that all essential details are preserved and that the output does not include any binary or base64 data. The output should be plain text without any formatting. Only provide the final text as the output: {resource}

# Apêndice C

## Instrução de Avaliação

Tabela C.1: Instrução de avaliação utilizando LLM como julgador

<b>Instrução de avaliação para comparação head-to-head</b>
<p>—Role—            You are an expert tasked with evaluating two answers to the same question based on four criteria: <b>Comprehensiveness</b>, <b>Diversity</b>, <b>Empowerment</b>, and <b>Directness</b>.</p> <p>—Goal—            You will evaluate two answers to the same question based on these four criteria:</p> <ul style="list-style-type: none"> <li>- <b>Comprehensiveness</b>: How much detail does the answer provide to cover all aspects of the question?</li> <li>- <b>Diversity</b>: How varied and rich is the answer in providing different perspectives and insights?</li> <li>- <b>Empowerment</b>: How well does the answer help the reader understand and make informed judgments about the topic?</li> <li>- <b>Directness</b>: How specifically and clearly does the answer address the question?</li> </ul> <p>For each criterion, choose the better answer (either Answer 1 or Answer 2) and explain why. Then, select an overall winner based on these four categories.</p> <p>Here is the question:            {query}</p> <p>Here are the two answers:  <b>Answer 1:</b>            {answer1}</p> <p><b>Answer 2:</b>            {answer2}</p> <p>Evaluate both answers using the four criteria listed above and provide detailed explanations for each.</p> <p>Output your evaluation in the following VALID JSON format:</p> <pre>           {{             "Comprehensiveness": {{               "Winner": "[Answer 1 or Answer 2]",               "Explanation": "[Provide explanation here]"             }},             "Diversity": {{               "Winner": "[Answer 1 or Answer 2]",               "Explanation": "[Provide explanation here]"             }},             "Empowerment": {{               "Winner": "[Answer 1 or Answer 2]",               "Explanation": "[Provide explanation here]"             }},             "Directness": {{               "Winner": "[Answer 1 or Answer 2]",               "Explanation": "[Provide explanation here]"             }},             "Overall Winner": {{               "Winner": "[Answer 1 or Answer 2]",               "Explanation": "[Summarize why this answer is the overall winner based on the four criteria]"             }}           }}           </pre>

# Apêndice D

## Principais Parâmetros

Listing D.1: Principais parametros para os experimentos

```
1 # LLM Configs default for Langchain
2 TEMPERATURE = 0.7
3 TOP_P = None
4 MAX_TOKENS = None
5
6 ## For Gemini model
7 safety_settings = {
8     HarmCategory.HARM_CATEGORY_UNSPECIFIED: HarmBlockThreshold.
9         BLOCK_NONE,
10    HarmCategory.HARM_CATEGORY_DANGEROUS_CONTENT:
11        HarmBlockThreshold.BLOCK_NONE,
12    HarmCategory.HARM_CATEGORY_HATE_SPEECH: HarmBlockThreshold.
13        BLOCK_NONE,
14    HarmCategory.HARM_CATEGORY_HARASSMENT: HarmBlockThreshold.
15        BLOCK_NONE,
16    HarmCategory.HARM_CATEGORY_SEXUALLY_EXPLICIT:
17        HarmBlockThreshold.BLOCK_NONE,
18 }
19
20 # Similarity Search configs on the information extraction tool
21 SIMILARITY_THRESHOLD = 0.84
22 K = 100000 # Number of number of retrival docs, practical non
23             limit.
24
25 # Embedding model
```

```
21 embedding_model = "BAAI/bge-small-en-v1.5"
22
23 # Database indexes
24 OPTIONS { indexConfig: {
25     'vector.dimensions': 384,
26     'vector.similarity_function': 'cosine'
27     }
28 }
```

# Apêndice E

## Exemplo de arquivo FHIR

Listing E.1: Exemplo do começo de um arquivo FHIR R4 em JSON para a paciente Beatris270 Bogan287

```
1 {
2   "resourceType": "Bundle",
3   "type": "transaction",
4   "entry": [
5     {
6       "fullUrl": "urn:uuid:5b3645de-a2d0-d016-0839-bab3757c4c58",
7       "resource": {
8         "resourceType": "Patient",
9         "id": "5b3645de-a2d0-d016-0839-bab3757c4c58",
10        "meta": {
11          "profile": [
12            "http://hl7.org/fhir/us/core/StructureDefinition/us-
13              core-patient"
14          ]
15        },
16        "text": {
17          "status": "generated",
18          "div": "<div xmlns=\"http://www.w3.org/1999/xhtml\">
19            Generated by <a href=\"https://github.com/
20              synthetichealth/synthea\">Synthea</a>. Version
21            identifier: 36ca5da\n .   Person seed:
22            -7838029442944994561   Population seed:
23            1699997975546</div>"
24        },
25        "extension": [
```

```

20     {
21         "url": "http://hl7.org/fhir/us/core/
           StructureDefinition/us-core-race",
22         "extension": [
23             {
24                 "url": "ombCategory",
25                 "valueCoding": {
26                     "system": "urn:oid:2.16.840.1.113883.6.238",
27                     "code": "2106-3",
28                     "display": "White"
29                 }
30             },
31             {
32                 "url": "text",
33                 "valueString": "White"
34             }
35         ]
36     },
37     {
38         "url": "http://hl7.org/fhir/us/core/
           StructureDefinition/us-core-ethnicity",
39         "extension": [
40             {
41                 "url": "ombCategory",
42                 "valueCoding": {
43                     "system": "urn:oid:2.16.840.1.113883.6.238",
44                     "code": "2135-2",
45                     "display": "Hispanic or Latino"
46                 }
47             },
48             {
49                 "url": "text",
50                 "valueString": "Hispanic or Latino"
51             }
52         ]
53     },
54     {
55         "url": "http://hl7.org/fhir/StructureDefinition/
           patient-mothersMaidenName",
56         "valueString": "Polly738 Lindgren255"

```

```

57     },
58     {
59         "url": "http://hl7.org/fhir/us/core/
           StructureDefinition/us-core-birthsex",
60         "valueCode": "F"
61     },
62     {
63         "url": "http://hl7.org/fhir/StructureDefinition/
           patient-birthPlace",
64         "valueAddress": {
65             "city": "Cambridge",
66             "state": "Massachusetts",
67             "country": "US"
68         }
69     },
70     {
71         "url": "http://synthetichealth.github.io/synthea/
           disability-adjusted-life-years",
72         "valueDecimal": 0.13231128504228493
73     },
74     {
75         "url": "http://synthetichealth.github.io/synthea/
           quality-adjusted-life-years",
76         "valueDecimal": 6.8676887149577155
77     }
78 ],
79 "identifier": [
80     {
81         "system": "https://github.com/synthetichealth/synthea
           ",
82         "value": "5b3645de-a2d0-d016-0839-bab3757c4c58"
83     },
84     {
85         "type": {
86             "coding": [
87                 {
88                     "system": "http://terminology.hl7.org/
                       CodeSystem/v2-0203",
89                     "code": "MR",
90                     "display": "Medical Record Number"

```

```

91         }
92     ],
93     "text": "Medical Record Number"
94 },
95     "system": "http://hospital.smarthealthit.org",
96     "value": "5b3645de-a2d0-d016-0839-bab3757c4c58"
97 },
98 {
99     "type": {
100         "coding": [
101             {
102                 "system": "http://terminology.hl7.org/
103                     CodeSystem/v2-0203",
104                 "code": "SS",
105                 "display": "Social Security Number"
106             }
107         ],
108         "text": "Social Security Number"
109     },
110     "system": "http://hl7.org/fhir/sid/us-ssn",
111     "value": "999-52-6108"
112 },
113 "name": [
114     {
115         "use": "official",
116         "family": "Bogan287",
117         "given": [
118             "Beatris270"
119         ]
120     }
121 ],
122 "telecom": [
123     {
124         "system": "phone",
125         "value": "555-339-8605",
126         "use": "home"
127     }
128 ],
129 "gender": "female",

```

```

130     "birthDate": "2015-05-03",
131     "address": [
132       {
133         "extension": [
134           {
135             "url": "http://hl7.org/fhir/StructureDefinition/
136               geolocation",
137             "extension": [
138               {
139                 "url": "latitude",
140                 "valueDecimal": 42.755478874424206
141               },
142               {
143                 "url": "longitude",
144                 "valueDecimal": -71.16438824378747
145               }
146             ]
147           },
148           "line": [
149             "274 Dooley Avenue"
150           ],
151           "city": "Lawrence",
152           "state": "MA",
153           "postalCode": "01843",
154           "country": "US"
155         }
156       ],
157       "maritalStatus": {
158         "coding": [
159           {
160             "system": "http://terminology.hl7.org/CodeSystem/v3
161               -MaritalStatus",
162             "code": "S",
163             "display": "Never Married"
164           }
165         ],
166         "text": "Never Married"
167       },
168       "multipleBirthBoolean": false,

```

```

168     "communication": [
169         {
170             "language": {
171                 "coding": [
172                     {
173                         "system": "urn:ietf:bcp:47",
174                         "code": "en-US",
175                         "display": "English (United States)"
176                     }
177                 ],
178                 "text": "English (United States)"
179             }
180         }
181     ]
182 },
183 "request": {
184     "method": "POST",
185     "url": "Patient"
186 }
187 },
188 {
189     "fullUrl": "urn:uuid:b3b4d7c6-46f5-9f10-cfd2-464ff96160f9",
190     "resource": {
191         "resourceType": "Encounter",
192         "id": "b3b4d7c6-46f5-9f10-cfd2-464ff96160f9",
193         "meta": {
194             "profile": [
195                 "http://hl7.org/fhir/us/core/StructureDefinition/us-
196                 core-encounter"
197             ]
198         },
199         "identifier": [
200             {
201                 "use": "official",
202                 "system": "https://github.com/synthetichealth/synthea",
203                 "value": "b3b4d7c6-46f5-9f10-cfd2-464ff96160f9"
204             }
205         ],
206         "status": "finished",

```

```

206     "class": {
207         "system": "http://terminology.hl7.org/CodeSystem/v3-
           ActCode",
208         "code": "AMB"
209     },
210     "type": [
211         {
212             "coding": [
213                 {
214                     "system": "http://snomed.info/sct",
215                     "code": "410620009",
216                     "display": "Well child visit (procedure)"
217                 }
218             ],
219             "text": "Well child visit (procedure)"
220         }
221     ],
222     "subject": {
223         "reference": "urn:uuid:5b3645de-a2d0-d016-0839-
           bab3757c4c58",
224         "display": "Beatris270 Bogan287"
225     },
226     "participant": [
227         {
228             "type": [
229                 {
230                     "coding": [
231                         {
232                             "system": "http://terminology.hl7.org/
           CodeSystem/v3-ParticipationType",
233                             "code": "PPRF",
234                             "display": "primary performer"
235                         }
236                     ],
237                     "text": "primary performer"
238                 }
239             ],
240             "period": {
241                 "start": "2015-05-03T13:37:42+00:00",
242                 "end": "2015-05-03T13:52:42+00:00"

```

```

243     },
244     "individual": {
245         "reference": "Practitioner?identifier=http://hl7.
                org/fhir/sid/us-npi|9999950899",
246         "display": "Dr. Elisa944 Rojo930"
247     }
248 }
249 ],
250 "period": {
251     "start": "2015-05-03T13:37:42+00:00",
252     "end": "2015-05-03T13:52:42+00:00"
253 },
254 "location": [
255     {
256         "location": {
257             "reference": "Location?identifier=https://github.
                com/synthetichealth/synthea|97876511-009c-391a-
                a262-1557e46861bb",
258             "display": "BLUESKIES WELLNESS INC"
259         }
260     }
261 ],
262 "serviceProvider": {
263     "reference": "Organization?identifier=https://github.
                com/synthetichealth/synthea|6ea4d3d6-d820-387a-a3d1-
                ed5944ad47a2",
264     "display": "BLUESKIES WELLNESS INC"
265 }
266 },
267 "request": {
268     "method": "POST",
269     "url": "Encounter"
270 }
271 }
272 ...
273 }

```

# Apêndice F

## Resultados por modelo

Tabela F.1: Comparação de estratégias de busca por modelo na escolha proporcional.

Métrica	Busca	Llama-3-70b	Gemini-1.5	Claude-v3	GPT-4o	DeepSeek-R1
<b>COMP</b>	L0	0.217	0.219	0.253	0.184	0.236
	L1	<b>0.328</b>	<b>0.359</b>	<b>0.353</b>	<b>0.379</b>	<b>0.352</b>
	S0	0.311	0.160	0.163	0.181	0.236
	S1	0.145	0.262	0.232	0.255	0.176
<b>DIV</b>	L0	0.204	0.280	0.211	0.192	0.245
	L1	<b>0.323</b>	0.263	<b>0.295</b>	<b>0.374</b>	<b>0.326</b>
	S0	0.264	0.165	0.211	0.142	0.176
	S1	0.209	<b>0.292</b>	0.284	0.292	0.253
<b>EMP</b>	L0	0.230	0.287	0.247	0.184	0.262
	L1	0.289	0.219	<b>0.274</b>	<b>0.358</b>	<b>0.322</b>
	S0	<b>0.294</b>	0.152	0.216	0.181	0.219
	S1	0.187	<b>0.342</b>	0.263	0.277	0.197
<b>DIR</b>	L0	<b>0.306</b>	<b>0.314</b>	<b>0.300</b>	<b>0.298</b>	<b>0.343</b>
	L1	0.281	0.236	0.253	0.248	0.240
	S0	0.213	0.205	0.200	0.223	0.189
	S1	0.200	0.245	0.247	0.230	0.227
<b>OW</b>	L0	0.217	0.274	0.258	0.184	0.245
	L1	<b>0.323</b>	0.266	<b>0.311</b>	<b>0.383</b>	<b>0.348</b>
	S0	0.302	0.143	0.142	0.163	0.227
	S1	0.157	<b>0.316</b>	0.289	0.270	0.180

Tabela F.2: Comparação de métricas ARS, CRS e HALS por modelo e estratégia de busca

Métrica	Busca	DeepSeek-R1	Claude-v3	Gemini-1.5	GPT-4o	Llama-3-70b
<b>ARS</b>	L0	0.87±0.3	0.83±0.29	0.97±0.09	0.9±0.25	0.86±0.3
	L1	0.88±0.26	0.95±0.06	0.97±0.11	0.93±0.15	0.9±0.2
	S0	0.88±0.3	0.9±0.24	<b>1.0±0.01</b>	0.9±0.24	0.82±0.36
	S1	0.84±0.33	0.83±0.33	0.96±0.09	0.88±0.28	0.82±0.33
<b>CRS</b>	L0	0.69±0.32	0.64±0.24	0.74±0.36	0.68±0.31	0.66±0.33
	L1	0.65±0.31	0.73±0.26	<b>0.82±0.31</b>	0.69±0.3	0.6±0.3
	S0	0.64±0.3	0.64±0.23	0.64±0.37	0.58±0.31	0.62±0.27
	S1	0.62±0.27	0.7±0.24	0.72±0.32	0.62±0.28	0.52±0.26
<b>HALS</b>	L0	0.32±0.47	0.22±0.42	0.06±0.23	0.16±0.36	0.28±0.45
	L1	0.3±0.47	0.48±0.51	0.11±0.31	0.34±0.48	0.37±0.49
	S0	0.18±0.38	0.3±0.47	0.08±0.26	0.14±0.36	0.23±0.42
	S1	0.42±0.49	0.41±0.5	<b>0.0±0.0</b>	0.34±0.48	0.45±0.5